UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ARMA 2020

July 8-9, 2020 Valencia, Spain

3rd International Conference on Advanced Research Methods and Analytics

ADE FACULTAD DE ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS. UPV

Departamento de Economía y Ciencias Sociales

devstat

Congress UPV

3rd International Conference on Advanced Research Methods and Analytics (CARMA 2020)

The contents of this publication have been evaluated by the Program Committee according to the procedure described in the preface. More information at http://www.carmaconf.org/

Scientific Editors

Josep Domenech
María Rosalía Vicente

Publisher

2020, Editorial Universitat Politècnica de València
www.lalibreria.upv.es / Ref.: 6563_01_01_01

Cover design by Gaia Leandri

ISBN: 978-84-9048-832-4 (print version)
Print on-demand

DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11920

3rd International Conference on Advanced Research Methods and Analytics (CARMA 2020)

This book is licensed under a Creative Commons Atribution-NonCommercial-NonDetivates-4.0 International license

Editorial Universitat Politècnica de València http://ocs.editorial.upv.es/index.php/CARMA/CARMA2020

Preface

Josep Domenech [1], María Rosalía Vicente [2]
[1] Dept. Economics and Social Sciences, Universitat Politècnica de València, Spain. [2] Dept. Applied Economics, Universidad de Oviedo, Spain

Abstract

Research methods in economics and social sciences are evolving with the increasing availability of Internet and Big Data sources of information. As these sources, methods, and applications become more interdisciplinary, the 3rd International Conference on Advanced Research Methods and Analytics (CARMA) is an excellent forum for researchers and practitioners to exchange ideas and advances on how emerging research methods and sources are applied to different fields of social sciences as well as to discuss current and future challenges. This edition was celebrated virtually because of the COVID-19 outbreak.

Keywords: *Big Data sources, Web scraping Social media mining, Official Statistics, Internet Econometrics, Digital transformation, global society.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Preface to CARMA2020

This volume contains the selected papers of the Third International Conference on Advanced Research Methods and Analytics (CARMA 2020) virtually hosted by the Universitat Politècnica de València, Spain during 8 and 9 July 2020. Despite the COVID-19 outbreak, This third edition consolidates CARMA as a unique forum where Economics and Social Sciences research meets Internet and Big Data. CARMA provides researchers and practitioners with an ideal environment to exchange ideas and advances on how Internet and Big Data sources and methods contribute to overcome challenges in Economics and Social Sciences, as well as on the changes in the society because of the digital transformation.

The selection of the scientific program was directed by Maria Rosalia Vicente, who led an international team of 67 scientific committee members representing institutions worldwide. Following the call for papers, the conference received 94 paper submissions from all around the globe. All submissions were reviewed by the scientific committee members under a double blind review process. Finally, 47 papers were accepted for oral presentation during the conference. This represents an overall paper acceptance rate of 50%, ensuring a high quality scientific program. It covers a wide range of research topics in Internet and Big Data, including official statistics, web scraping, search engine data, industry adoption, sentiment analysis, geospatial data or consumer behavior, among others. Additionally, 12 posters with promising work-in-progress research were selected for presentation during the conference.

Apart from the regular scientific sessions, the keynote speech was contributed by Pablo de Pedraza, who talked about “The semicircular flow of the data economy” and provided a unique view from his position in the Monitoring, Indicators and Impact Evaluation Unit at the Joint Research Centre of the European Commission.

The conference organizing committee would like to thank all who made this third edition of CARMA a great success. Specifically, thanks are indebted to the authors, scientific committee members, reviewers, invited speaker, 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 2020.

2. Organizing Committee

General chair

Josep Domènech, Universitat Politècnica de València

Scientific committee chair

María Rosalía Vicente, Universidad de Oviedo

Local organization

Eduardo Cebrián
Mónica Costa Alcaina
Eduardo Torán

3. Sponsors and Supporters

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

4. Scientific committee

Anto Aasa, University of Tartu
Fernando Almeida, University of Porto & INESC TEC
Helena Alves, University of Beira Interior
María del Pilar Ángeles, Universidad Nacional Autónoma de México
Concha Artola, Banco de España
Nikolaos Askitas, IZA – Institute of Labor Economics
Seyhmus Baloglu, University of Nevada
Catherine Beaudry, Polytechnique Montreal
Silvia Biffignandi, University of Bergamo
Federico Botta, University of Exeter
Petter Bae Brandtzaeg, University of Oslo/SINTEF digital
Levent Bulut, Valdosta State University
Ludovic Calès, European Commission, JRC
José Luis Cervera, DevStat
Cihan Cobanoglu, University of South Florida
Marisol B. Correia, ESGHT-Algarve University & CiTUR
Piet J.H. Daas, Statistics Netherlands/Eindhoven University of Technology
Stefano De Marco, University of Salamanca

Pablo de Pedraza, European Commission, JRC
Giuditta de Prato, European Commission, JRC
Thomas Dimpfl, University of Tübingen
Carlo Drago, University Niccolò Cusano
Rameshwar Dubey, Montpellier Business School
Enrico Fabrizi, Università Cattolica del S. Cuore
Mohammad Falahat, Universiti Tunku Abdul Rahman
Juan Fernández de Guevara, Universitat de València & Ivie
Youssef Gahi, University of Ibn Tofail
Rui Gaspar, Universidade Catolica Portuguesa
Marcos González-Fernández, Universidad de León
Peter Hackl, Vienna University of Economics and Business
Abdul Hafeez, University of Engineering & Technology
Agustín Indaco, Carnegie Mellon University in Qatar
Zaheer Khan, University of the West of England
Jan Kinne, ZEW Centre for European Economic Research
Felix Krupar, IOTA Foundation
Diego Kuonen, Statoo Consulting & University of Geneva
Caterina Liberati, Università di Milano-Bicocca
Francisco Martínez-Álvarez, Pablo de Olavide University
Rocío Martínez-Torres, Universidad de Sevilla
Gavin McArdle, University College Dublin
Jesús Morán, University of Oviedo
Igor Mozetic, Jozef Stefan Institute
María Olmedilla, SKEMA Business School
Irem Onder, University of Massachusetts Amherst
Enrique Orduña-Malea, Universitat Politècnica de València
José Luis Ortega, Institute for Advanced Social Studies (IESA-CSIC)
Luca Pappalardo, ISTI-CNR
José Manuel Pavía Miralles, Universitat de València
Viktor Pekar, Aston University
Arturo Peralta Martín-Palomino, University of Castilla-La Mancha
Ricardo Pérez del Castillo, University of Castilla-La Mancha
Maria Petrescu, ICN Business School Artem, CEREFIGE Lab., France Colorado State University Global
Ana Pont, Universitat Politècnica de València
Bruce Prideaux, Central Queensland University
Ravichandra Rao, Indian Statistical Institute
Pilar Rey del Castillo, Instituto de Estudios Fiscales

Rosa Rio-Belver, Universidad del Pais Vasco
Anna Rosso, University of Milano DEMM
Pål Sundsøy, NBIM
Sergio Toral Marin, Universidad de Sevilla
Konstantinos P. Tsagarakis, Democritus University of Thrace
Joonas Tuhkuri, MIT
Tiziana Tuoto, Istat Italian National Institute for Statistics
Antonino Virgillito, Italian Revenue Agency
Maro Vlachopoulou, University of Macedonia/Greece
Martin R. Wolf, University of Applied Sciences Aachen
Selim Zaim, Istanbul Sehir University

Index

Full papers

A method for determining the emergence level of transformer technologies for green energy applications 1

Mining News Data for the Measurement and Prediction of Inflation Expectations 9

Citizens' attention in Madrid City through the study of personalized records 19

Investigating the impacts of street environment on pre-owned housing price in Shanghai using street-level images 29

eWOM in reward-based crowdfunding platforms: A behavioral approach 41

An algorithm to fit conditional tail expectation regression models for vehicle excess speed in driving data 51

Regression scores to identify risky drivers from braking pulses 59

Pruned Wasserstein Index Generation Model and wigpy Package 69

Model degradation in web derived text-based models 77

A field study on the impacts of implementing concepts and elements of industry 4.0 in the biopharmaceutical sector 85

High order PLS path modeling to evaluate well-being merging traditional and big data: A longitudinal study 95

Big Data in Corporate Governance decision 103

Question-Generating Datasets: Facilitating Data Transformation of Official Statistics for Broad Citizenry Decision-Making 113

Evaluating accredited mHealth applications. An exploratory study 123

Strategic Open Innovation model: Mapping Iberdrola network 133

Data granularity in mid-year life table construction 143

Extracting User Behavior at Electric Vehicle Charging Stations with Transformer Deep Learning Models 153

Comparative multivariate forecast performance for the G7 Stock Markets: VECM Models vs deep learning LSTM neural networks 163

Investigating inefficiencies of bookmaker odds in football using machine learning 173

Sentiment Analysis of Twitter in Tourism Destinations 181

Google Trends Topic-Based Uncertainty: A Multi-National Approach 191

Bridging internet and cultural heritage through a digital marketing funnel: An exploratory approach 201

Combining content analysis and neural networks to analyze discussion topics in online comments about organic food 211

Setting Crunchbase for Data Science: Preprocessing, Data Integration and Feature Engineering 221

Information balance between newspapers and social networks 231

Third Places and Art Spaces: Using Web Activity to Differentiate Cultural Dimensions of Entrepreneurship Across U.S. Regions 239

New technologies and role of direct surveys in the production of Official Statistics 247

Sample Size Sensitivity in Descriptive Baseball Statistics 253

Extracting usual service prices from public contracts 259

Communicating Corporate Social Responsibility through Twitter: a topic model analysis on selected companies 269

Proposal of a composite indicator for measuring social media presence in the wine market 279

Political Polarization and Movie Ratings: Web Scraping The Brazilian Contemporary Scenario 289

Comparing Methods to Retrieve Tweets: a Sentiment Approach 299

Donald Trump, investor attention and financial markets 307

#immigrants project: the on-line perception of integration 321

Abstracts

Digital footprint for tourism research 333

Predicting SME's default: some old facts and a new idea 334

Journalists as end-users: quality management principles applied to the design process of news automation 335

Identification of online reviews helpfulness using Neural Networks 336

User-defined Machine Learning Functions 337

Internet searches as a leading indicator of house purchases in a subnational framework: the case of Spain 338

Causal discovery with Point of Sales data 339

Interpretable Machine Learning - An Application Study Using the Munich Rent Index ... 340

Enhancing UX of analytics products with AI technology 341

Search in second Hand market : The case of mobile phone 342

The epistemological impacts of big data on public opinion studies 343

Measuring and Forecasting Job-Search in Italy using Machine Learning 344

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11315

A method for determining the emergence level of transformer technologies for green energy applications

Gaizka Garechana, Rosa Río-Belver, Enara Zarrabeitia, Izaskun Álvarez-Meaza
Department of business management, University of the Basque Country, Spain.

Abstract

Solid State Transformers (SST) are the result of merging the power electronics possibilities for voltage and frequency control with high-frequency transformers, and are expected to be a key component for enabling some important features that future energy grids must possess: reversibility, stability, modularity and compactness, among others. In addition to this, the possibilities of SSTs can be enhanced with advanced semiconductor materials such as Silicium Carbide (SiC), considerably improving the voltage and frequency ranges of this devices. This study aims at developing a quantitative method for characterizing the emergence level of SSTs and SiC-based transformers in three areas where these technologies can have a sizable impact: photovoltaic (PH) and eolic (EO) energy production and electric vehicle (EV) appliances. Results show that PH area will probably outpace the EO area in both technologies, but the attention of the scientific community may be shifting from PH in the SiC-based transformer technology. EV applications are, on average, closer to the life cycle's exponential growth stage than PH and EO areas, so it seems reasonable to expect a comparatively faster increase of both scientific and technology development activity in this field.

Keywords: *SST; Silicium Carbide; Technology Forecasting; Emergent technologies.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

Electrical transformers are devices that allow changing the voltage of electric current, a well-known application of transformers is that of increasing the voltage of the alternating current (AC) generated in power stations from low to high (step-up transformer), in order to increase the voltage and reduce the current, given that less current means that less energy is lost when transporting said current to the customers. High voltage current, however, is dangerous for typical household purposes, so voltage must be reduced back to safe levels (step-down transformer) before its final use. The conventional transformer presents some drawbacks for its deployment in local, decentralized, renewable energy grids. First, conventional transformers are too big for many of these applications. Second, transformers are one-way tools, suited for energy distribution systems designed around big, centralized power plants, which comes in stark contrast with the operational aspects of the smart and clean technologies that are expected to substantially increase their share in the energy mix of the future (Roberts, 2018). There is a need for small and flexible transformation systems that can also deal with energy storage systems that work on direct current (DC).

Solid State Transformers (SST) are the result of merging the power electronics possibilities for voltage and frequency control with high-frequency transformers. One of the core tasks of power electronics in these devices is to increase the typical current frequency coming from the grid (50 Hz in Europe) to a range between 10 and 20 KHz in order to feed a high-frequency transformer that could be 20% smaller than a conventional transformer. This can be achieved using conventional silicon-based insulated-gate bipolar transistors (IGBT), at the expense of the reliability and limited handling of voltage (6.5 Kv). New semiconductor materials such as the silicium carbide (SiC) address these shortcomings, enabling SSTs to work at higher voltages and very high frequencies, thus achieving the maximum reduction in transformer size (Bhattacharya, 2017). The flexibility brought by the power electronics also allows the adjustment of the SST to frequent shifts in voltage and the requirements of a smart energy-management system (Abu-Siada, Budiri, & Abdou, 2018).

The three-module approach proposed by Bhattacharya (2017) offers direct DC connection in the components of the SST, allowing to build direct interfaces with solar or other renewable energy technologies, thus avoiding extra DC – AC conversion steps and consequently, improving the efficiency. The possibilities of this multi-port structure go even further: Three module SiC-based SST systems can also be used to provide high voltage DC connection ports for electric vehicle (EV) quick chargers, using compact devices. The reversibility of the system could even allow using the local fleet of electric cars as a storage system for backing up the grid when necessary (Ronanki, Kelkar, & Williamson, 2019).

2. Research goals

The goal of the research presented in this paper is to develop a method suitable for determining the relative (since all the technologies studied in this paper are considered to be "emergent" according to the industry consensus) maturity of the applications of SST technology in the areas of photovoltaic energy (PH), eolic energy (EO) and the electric vehicle (EV), where the new transformer technologies are deemed to have a high impact.

The above mentioned method will also be applied to transformer technologies based on the advanced semiconductor material SiC, in order to analyze the penetration of this semiconductor in the transformer industry, in the PH, EO and EV areas.

3. Methodology

The basic premise underlying our methodology is that the emergence stage of the technology life cycle corresponds with the exponential growth stage of the logistic growth curve (Kucharavy & De Guio, 2015) and consequently, the current degree of development and future prospectives of an emergent technology can be characterized by fitting the data corresponding to that technology to an exponential model.

3.1. Data retrieval and subsetting

The present study uses scientific publication and patent data for the characterization of the developments taking place in a technological field. According to the linear model of innovation (Godin, 2006), advances in scientific activity should come before the development efforts (patents) at the organizations. Data was retrieved by running the following queries on Scopus database (scientific publications):

- SST technology: SST: TITLE-ABS-KEY (solid W/0 state W/0 transformer)
- SiC in transformer technology: (TITLE-ABS-KEY ("SILICON CARBIDE" OR "SILICIUM CARBIDE" OR carborundum OR sic) AND TITLE-ABS-KEY (transformer))

…and their approximate equivalents in Patseer (patents) database:

- SST technology: (TACD:(SOLID wd0 STATE WD0 TRANSFORMER))
- SiC in transformer technology: ((TACD:(("silicon carbide" OR "silicium carbide" OR "carborundum") AND TRANSFORMER) AND AC:(H02M*)))

In order to accomplish the goals stated in section 0, data was subsetted by applying text mining techniques to the "description" field of patents and "title, abstract & keywords" fields of scientific publications. We assume that the presence of terms unequivocally associated with PH, EO or EV applications in these fields is an evidence that can be used for subsetting

our data in three subsamples corresponding to the aforementioned areas. This approach provides a total amount of twelve subsamples to perform our study: six subsamples corresponding to scientific activity (PH, EO, EV – Scopus) in both SST and SiC transformer technologies, and their equivalent in patenting activity, (PH, EO, EV – Patseer) in both SST and SiC transformer technology.

These being emergent technologies, we obtained very few data prior to year 2000, so the time interval was defined from 2003 to 2017 for SiC transformer data, and from 2011/12 to 2018 for SST data. We pragmatically selected the start of the interval by selecting the first year in which at least one record was obtained for two out of the three areas under analysis (PH, EO, EV), for both SST and SiC transformer technologies.

3.2. Fitting the data to the exponential curve

The number of publications and patents corresponding to the twelve subsamples is fitted to an exponential model, according to the following method:

First, we parametrize the equation $y = ax^b$ by taking logs $\log(y) = \log(a) + b * \log(x)$ and fitting a linear regression model to the log-transformed data . This will provide an initial estimation of a and b parameters that will be subsequently recalculated by using the nonlinear weighted least-squares (NLS) method on the data, as explained in Hastie (2017), using R. The total amount of patents and publications corresponding to each area (PH, EO, EV) will inform abut the diffusion achieved by the technologies under study (SST and SiC based transformers) while the parameter b will inform about the relative maturity of said technologies in each area, considering that the higher the b, the higher are the expectations about the success of the technology in a particular area, and the closer is that technology to the exponential growth phase of the technology life cycle.

4. Results

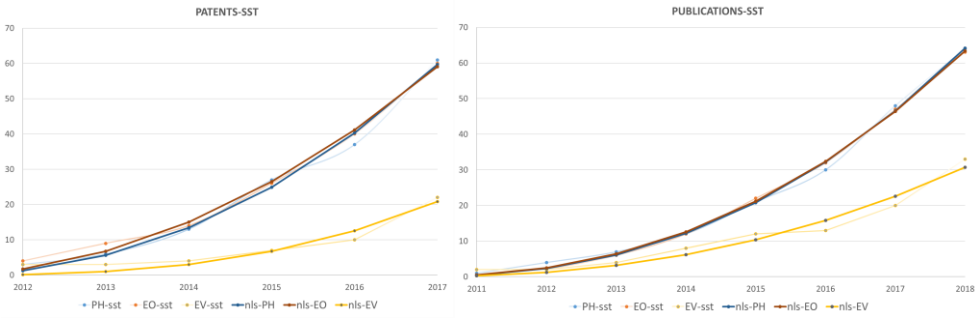


Figure 1 shows the accumulated patent and publication data corresponding to SST technology, as well as the results of fitting the model to the data:

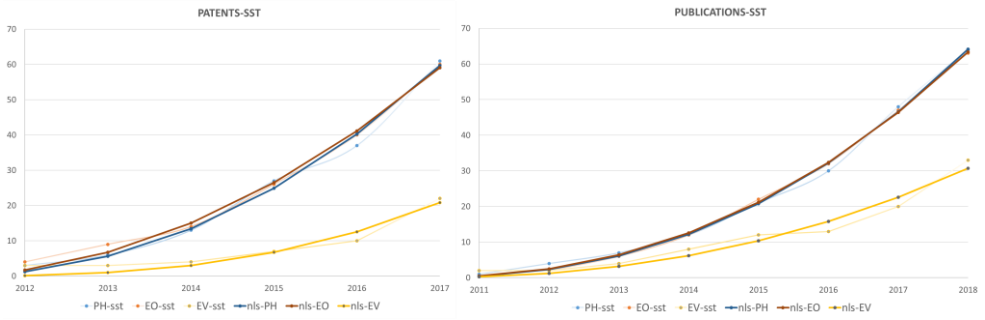


Figure 1. Data corresponding to SST technology (accumulated publications and patents). Colors indicate the application area (PH, EO, EV) and the thickness of the line indicates whether it shows raw data (thin) or the result of fitting the NLS model (thick).

The first thing we notice is that both data sources (publications and patents) show approximately the same starting point for the data, according to the criteria we exposed on section 0, and the patterns shown by data are also similar for both sources. The applications of SST to the EV area are fewer in number, and start to show a growth pattern later, when compared with PH and EO data. However, the *b* parameter is significantly higher for patent data in EV area, as can be seen in the results presented in Table 1.

Figure 2 shows the results corresponding to the SiC transformer technology:

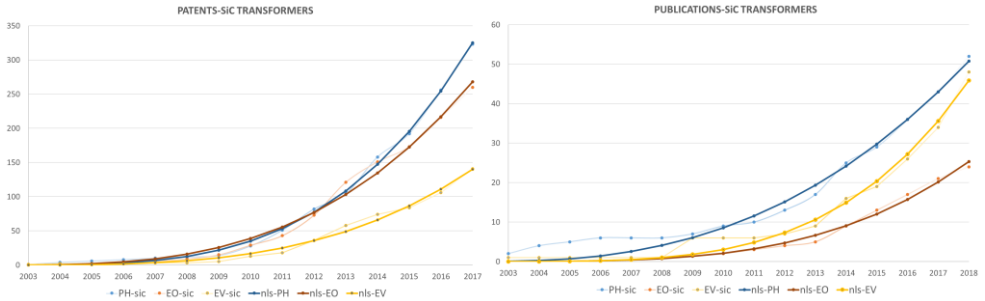


Figure 2. Data corresponding to SiC transformer technology(accumulated publications and patents). Colors indicate the application area (PH, EO, EV) and the thickness of the line indicates whether it shows raw data (thin) or the result of fitting the NLS model (thick).

We observe the same phenomenon regarding the starting point of time intervals: both patent and publication data corresponding to this technology show the same starting point, according to our methodology. In this case, however, the pattern shown by the areas under analysis is significantly different for each data source. While the EV applications remain below PH and EO in patent data, research in EV applications is clearly outpacing EO and

shows signs of surpassing the PH area, as can be seen in the *b* parameter values shown in Table 1. EO and EV areas seem to be catching the attention of the scientific community at a fast pace when compared with PH, by looking at their *b* values. This was not the case in SST technology. However, PH applications of SiC based-transformers are the most frequent, according to our data.

Table 1 shows the values of *b* for the NLS models built for each technology, area and data source:

Table 1. Values of b for each technology, area and data source.

Technology - Area	Data source	*b* parameter
SST-PH	Development (patents)	2.15
SST-EO	Development (patents)	1.97
SST-EV	Development (patents)	2.79
SST-PH	Research (scientific pub.)	2.39
SST-EO	Research (scientific pub.)	2.33
SST-EV	Research (scientific pub.)	2.31
SIC-PH	Development (patents)	3.54
SIC-EO	Development (patents)	3.07
SIC-EV	Development (patents)	3.38
SIC-PH	Research (scientific pub.)	2.57
SIC-EO	Research (scientific pub.)	3.58
SIC-EV	Research (scientific pub.)	3.91

The results show that EV applications have the highest average *b* coefficient (3.09), while PH and EO applications show similar *b* values (2.66 and 2.73, respectively).

5. Discussion and conclusions

A remarkable conclusion of this study is that the ideas of the linear model of innovation (Godin, 2006) fail to describe the behavior of SST and SiC transformer technologies: both research and development seem to be taking place simultaneously, according to our data. Research in these technologies is eminently applied science, where the boundaries between science and development become more porous (Kline, 1985), this could be an explanation of

the phenomenon we have observed, but for our purposes, this trait of the technologies hinders the detection of early signs of emergence coming from the scientific world.

A head to head comparation between the two renewable energy sources under study suggests that the PH area will probably outpace the EO area in both technologies (SST and SiC-based transformers). The trends in scientific research in SiC-based transformers, however, suggest that the attention of the scientific community might be shifting from PH in this technology.

Perhaps the most interesting pattern can be found in the EV applications of both SST and SiC transformers. Data corresponding to EV area persistently shows a smaller yearly amount of research/patenting activity taking place when compared with the rest of the areas (with a single exception) but at the same time the higher average *b* parameter is found in this area. According to our approach, this suggests that EV applications might be closer to the life cycle's exponential growth stage than PH and EO areas, so it seems reasonable to expect a comparatively faster development of both scientific and development activity in this field, when compared with PH and EO. This conclusion is reinforced, from our point of view, by the data presented in Figure 2 (right), which points at both a strong presence and exponential growth pattern in the academic activity related to SiC transformer applications in EV area. Considering that a substantial amount of research related with new semiconductor materials falls into the realm of basic science, technology forecasting efforts in this area should probably keep an eye on this sample of data, in order to look for early signals of emergence.

We hope that the results and the conclusions presented in this study will be useful for decision making in the field of renewable energies and the technologies related to electric vehicles, particularly for those professionals involved in technology forecasting practices.

References

Abu-Siada, A., Budiri, J., & Abdou, A. (2018). Solid State Transformers Topologies, Controllers, and Applications: State-of-the-Art Literature Review. *Electronics*, *7*(11), 298. https://doi.org/10.3390/electronics7110298

Bhattacharya, S. (2017). Smart Transformers Will Make the Grid Cleaner and More Flexible. Retrieved January 21, 2020, from https://spectrum.ieee.org/energy/renewables/smart-transformers-will-make-the-grid-cleaner-and-more-flexible

Godin, B. (2006). The Linear Model of Innovation: The Historical Construction of an Analytical Framework. *Science, Technology & Human Values*, *31*(6), 639–667. https://doi.org/10.1177/0162243906291865

Hastie, T. J. (2017). *Statistical models in S*. (T. J. Hastie & J. M. Chambers, Eds.). Routledge.

Kline, S. (1985). Innovation is not a linear process. *Research Management*. Retrieved from http://www.ec.unipg.it/ez_new/index.php/ita/content/download/7711/35914/file/FILE 3_ Kline_ Innovation is not a linear process.pdf

Kucharavy, D., & De Guio, R. (2015). Application of Logistic Growth Curve. *Procedia Engineering*, *131*, 280–290. https://doi.org/10.1016/J.PROENG.2015.12.390

Roberts, D. (2018). Renewable energy threatens to overwhelm the grid. Here's how it can adapt. Retrieved January 21, 2020, from https://www.vox.com/energy-and-environment/2018/11/30/17868620/renewable-energy-power-grid-architecture

Ronanki, D., Kelkar, A., & Williamson, S. S. (2019). Extreme Fast Charging Technology—Prospects to Enhance Sustainable Electric Transportation. *Energies*, *12*(19), 3721. https://doi.org/10.3390/en12193721

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11322

Mining News Data for the Measurement and Prediction of Inflation Expectations

Diana Gabrielyan[1], Jaan Masso[1], Lenno Uusküla[2]
[1]University of Tartu, Tartu, Estonia, [2]Bank of Estonia, Tallinn, Estonia.

Abstract

In this paper we use high frequency multidimensional textual news data and propose an index of inflation news. We utilize the power of text mining and its ability to convert large collections of text from unstructured to structured form for in-depth quantitative analysis of online news data. The significant relationship between the household's infla-tion expectations and news topics is documented and the forecasting performance of news-based indices is evaluated for different horizons and model variations. Results sug-gest that with optimal number of topics a machine learning model is able to forecast the inflation expectations with greater accuracy than the simple autoregressive models. Addi-tional results from forecasting headline inflation indicate that the overall forecasting accu-racy is at a good level. Findings in this paper support the view in the literature that the news are good indicators of inflation and are able to capture inflation expectations well.

Keywords: *inflation; inflation expectations; news data; natural language processing; topic modelling.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Household surveys of inflation often indicate that the perception of the current inflation differs substantially from the actual values of inflation. Similarly, expectations about the future expectations differ strongly from the surveys of professional forecasters and the implied inflation rates of financial markets (for evidence see e-g- Coibion et al. 2018). Potential reason for the difference is that households and firms obtain only very partial information while doing everyday purchases and aggregating the information is very costly. Imperfect information in turn affects adversely the formation of expectations. Subjective inflation nowcasts and expectations are built through personal experiences, prior memories of inflation, and various other sources of information. One primary source of information is public media and it is well established that consumers rely largely on it when thinking about overall price changes (Blinder and Krueger 2004, Curtin 2007). Media covers a lot of news on prices and price developments.

In this paper we explore online news as novel data source for measuring inflation perception and forecasting inflation expectations by utilizing the power of text mining and its ability to convert large collections of text from unstructured to structured form. We propose a novel index of inflation news that provides a real-time indication of the price developments. Such index of inflation news captures and summarizes well the information used in the formation of expectations[1]. Available survey-based inflation expectations have low frequency and the high-frequency market-based forecasts involve risk premia and may be uncertain[2]. Our main contribution is therefore using the novel source of information to prove that online news can provide a real-time and accurate indication of consumer's expectations on inflation.

Machine learning methods are considered to be very promising avenue for academic and applied research. Although its applications are already actively used in many disciplines and research areas, it is still relatively new to economics. One modern strand of machine learning is text mining – the computational approach to processing and summarizing large amounts of text, which would be far more difficult to read, even impossible, for any single person. Extracting information from novel sources of data, such as social media (e.g. Twitter, Google) or public media (e.g. online news, communication reports) allows analysis and

[1] As Nimark and Pitschner (2018) note, since no agent has resources to monitor all events potentially relevant for his decisions, news are preferred delegates for information choice to monitor the world on their behalf. And since news mainly reports selection of events, typically major ones, coverage becomes more homogenous across different outlets.

[2] Market-based expectations are available daily but include risk premia. Survey-based expectations are published monthly. For example, for the United Kingdom, the quarterly Consumer trends data are typically published around 90 days after the end of the quarter. See https://www.ons.gov.uk/economy/nationalaccounts/satelliteaccounts/bulletins/consumertrends/apriltojune2019

different kind of understanding of economics relationships, e.g. consumer behaviour, therefore contributing to policy making and forecasting. See for example, Tuhkuri (2016), D'Amuri and Marcucci (2017), Yu et al (2018), Nyman et al (2015).

Another contribution of this work is to forecast the inflation in real-time using machine learning methods. The importance of inflation forecasting for rational decision making is well established in the literature along with the common knowledge that improving upon simple models is quite challenging. According to Medeiros et al (2019), most of this literature however ignores the recent machine learning advances. In their work they show that with machine leaning and data-rich models improving inflation forecasts is possible. Their LASSO and Random Forest models are able to produce more accurate forecasts than the standard benchmark models, e.g. autoregressive models. Similarly, Garcia, Medeiros and Vasconcelos (2017) find that high dimensional models perform very well in inflation forecasting in data rich environments. Our findings from LASSO regressions support these findings: for inflation expectations the short-term forecast errors are smaller than those of the autoregressive models. The analysis also identifies the optimal number of news topics for predicting up to five quarters ahead inflation expectations to be either four or five suggesting that the LASSO regression using optimal number of topics and best value of regularization parameter results in simpler model, which doesn't compromise the model performance. These results are, however, not robust for longer forecasting horizons and for different values of the regularization parameter. In additional results, when forecasting headline inflation, we find that the LASSO models fail to improve upon the benchmark models but demonstrate similar forecasting accuracy.

The rest of the paper is organized as follows. Section 2 describes the data sources and methodology. Section 3 and 4 provide results and an application in forecasting respectively. Section 5 concludes.

2. Data and Methodology

For official statistics, we use the Bank of England Inflation Attitude Survey data and actual UK inflation statstics. Our novel inflation news indicator is built from the article data of one of the UK leading newspaper's , Guardian, business section over the last 15 years. The choice of the news outlet is due relevance to our research in terms of content and readership, as well as the availability of open source data. As such, we chose Guardian news data for our analysis. Any news in Guardian is public and readable by anyone by default. Overall, we collected around 20,000 documents and 32 million terms from January 2004 to January 2019, which is sufficient amount of data to conduct our analysis. We only fetch articles from the business section, since this is the most relevant section for economic topics in general. In addition, articles were also filtered based on subjectively chosen key-words, which in our

opinion are relevant to inflation expectations topic. Namely, they are price, price increase, expensive, cheaper, cost, expense, bill, payment, oil, petrol, gas, diesel. The data comes in unstructured form, that is, the data is in a text form and does not have a given structure. Overall, our news corpus consists of around 100000 English language articles with well above 20 million words from January 2004 to January 2019, which is sufficient amount of data to conduct our analysis. However, the amount of data, also makes statistical computations challenge. We therefore apply data pre-processing steps suggested by Bholat and co-authors (2015) at the same time adding more steps and more developed methods. We use the text mining's bag of word approach in the text, which means all words are analysed as a single token and their structure, grammar or part of lexicon does not matter. Pre-processing results in a document term matrix, which includes all occurrences of the words in the corpus and their respective frequencies. At this step, the dimensionality of the corpus is reduced, and we get more understandable results. Frequency counts of the top 31 words in their stemmed form, that is the number of times those words appear in the final sample, are plotted in Figure 1.

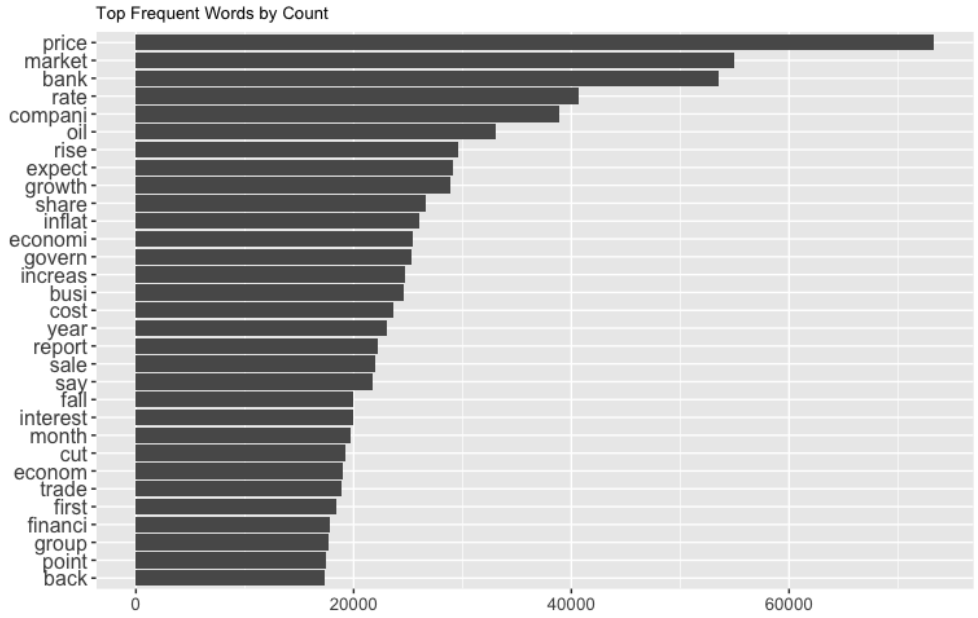


Figure 1. Top frequent words and their counts. The words are presented in stemmed form.

To proceed to building the index, we proceed with topic modelling. Since any document can be assigned to several topics at a time, the probability distribution across topics for each

document is therefore needed. Latent Dirichlet Allocation (LDA) [3] is a statistical model that identifies each document as a mixture of topics (related to multiple topics) and attributes each word to one of the document's topics, therefore, clustering words into topics. With LDA method, it is possible to derive their probability distribution by assigning probabilities to each word and document. Assigning words and documents to multiple topics also has the advantage of semantic flexibility (ex. the word 'rate' can relate both to inflation and unemployment topic). Thorstrud (2018) notes that LDA shares many features with Gaussian factor models, with the difference being that factors here are topics and are fed through a multinomial likelihood. In LDA, each document is given a probability distribution and for each word in each document, a topic assignment is made.

3. Results

For each document within a day, five most popular words are identified, and their daily frequency is calculated. This allows counting also the frequency of each topic for a given day. At this step, our results of topic decompositions and distribution are used to build the new high frequency index that will capture the intensity of inflation expectations. The index is built for every day, that is, we build daily time series using Guardian's business articles for each day. To do so, we first collect together all articles for a given day into one document, grouping them into one plain text for each day. Next, based on the first ten most frequent words in each topic the article's daily frequency is calculated. In other words, the frequency is calculated for the given day as the raw count of frequencies with which the most common words in each topic appear in that day. The news volume *I(t)* of given topic *z* is given by

$$I_z(\mathrm{t}) = \sum_{d \in I(t)} \sum_w N(d, w, z), \tag{2}$$

where $N(d, w, z)$ is the frequency with which the word *w* tagged with topic z appears in document d. These time series $I_z(\mathrm{t})$ are measures of volume, that is, they measure the intensity of given topic for given time period, that is for given day.

We find that some of index series are non-stationary and consequently transform them to stationary series by differencing. Augmented Dickey Fuller test is used to determine the presence of unit root and hence understand if the series are stationary or not. As such, some of the indices are evaluated as non-stationary and are transformed to by differencing.

[3] Detailed description of the LDA approach is provided in Blei, Ng and Jordan (2003).

4. Application in Forecasting

The first task is to filter information from the list of variables and select more relevant components. It is highly inefficient to use all the topic indices for predicting in such a rich dataset, as some of the regressors may be imparting redundant information. Therefore, number of topics *N* is too high and there is a definite multicollinearity present among the topic indices, as can also be observed from Figure 4. To reduce dimensionality and tackle the issue of multicollinearity[4], we use another machine learning method for variable selection. LASSO (Least Absolute Shrinkage and Selection Operator) method automates variable selection by reducing the coefficients of some features to zero, while keeping those that have the most impact on the dependent variable. LASSO's main goal is finding β that minimizes (3) with constraint $\sum_{j=1}^{p}|\beta_i| \leq t$.

$$\sum_{i=1}^{M} (y_i - \sum_{j=1}^{p} \beta_i x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_i| \quad (3)$$

λ is the shrinkage parameter and controls the strength of penalty finding the model with the smallest number of predictors that also gives a good accuracy. Therefore, the number of variables to be removed is decided by the shrinkage parameter λ, which is chosen using cross validation. Once the topic indices are selected, we forecast the inflation expectations by building a model using a direct forecast approach as given in equation (4).

$$\pi_t^{t+h} = \alpha + a * \pi_{t-1}^{t-1+h-1} + \sum_{n=1}^{N} b_n * x_{n,t-1} + u_t, \quad (4)$$

where π_t^{t+h} is the inflation (expectations) for the next *h* quarters at period *t* and $\pi_{t-1}^{t-1+h-1}$ the lagged value for the same horizon. *N* is the number of indices built from news data, b_n are vectors of unknown parameters, $x_{n,t}$ are the lagged indices and u_t is the forecasting error. We call the Equation (3) a news-based model (NBM). It is common practice to fit a model using training data, and then to evaluate its performance on a test data set. Forecast horizon *h* is also the length of the out-of-sample period (i.e. fitted values on the training set) and will be varied from 1 to 12 to compare the forecasts at different horizons and find the 'optimal' horizon defined by the lowest forecasting error. Since all of the data in this analysis is quarterly, *h* is measured in quarters. For benchmarking we use naïve AR (1) model on inflation expectations and compare the root mean squared errors (RMSE).

Table 1 reports the normalized results of estimating (3) and an AR (2) with different forecast horizons relative to simple AR (1) model. The first column of the table shows the forecast horizon, the second column (*n_var*) shows the number of variables (topics) selected by LASSO regression, and the last two columns show the root mean squared errors (RMSE) for each of the applied models. It can be seen that generally, the RMSEs are small, varying from

[4] LASSO is very robust against multicollinearity, see Friedmanet al. (2001).

0.02 to 0.76), while the forecast errors are the lowest when forecasting the next one or two period expectations using the news data. In this case the LASSO model outperforms both the naïve AR (1) and AR (2) forecasts in terms of accuracy.

Table 1. RMSEs of h-period inflation expectations forecasts using LASSO and AR (2) models. Errors are normalized relative to AR (1) benchmark.

h	n_min	RMSE_LASSO_MIN	RMSE_AR2
1	5	0.6	6
2	5	0.7	1.9
3	6	0.9	1.8
4	5	0.8	1.8
5	5	0.8	1.8
6	5	1	1.6
7	5	1.1	1.7
8	5	1	2.1
9	5	1.2	2.9
10	4	1.6	1.9
11	3	1.3	1.8
12	3	1.5	1.8

Several interesting observations can be made from Table 1. Firstly, LASSO models select different number of topics that are relevant for inflation expectations prediction for different forecast horizons. Out of our fifty topics compiled by the LDA method, LASSO selects three to six topics depending on the forecast horizon. Lagged value of the inflation expectations is always included among selected regressors and is always significant. The adjusted R-squared statistic is informative and for some horizons is as high as 70%. Thus, the selected news topic, as well as the past values of inflation expectations explain a relatively large fraction of the variation in the household's inflation expectations. One to two quarters ahead expectations can be forecasted with five topic indices as regressors, while the longer forecasts of eleven and twelve quarters can be forecasted with the best accuracy when only three

relevant topics are employed in the regression. It can also be observed that the longer the forecast horizon, the lower the forecast accuracy, which is intuitive.

These results were not robust when controlling and comparing different values of regularization parameter in the LASSO regression. There are different ways to choose the optimal value of lambda by cross-validation. Our main results in Table 1, where based on the smallest value of lambda from the cross-validation results. Table 2 compares the accuracy obtained with LASSO regression using different values of lambda shrinkage parameters against the benchmark autoregressive models. First column is the forecast horizon, while following 3 columns report the number of regressors selected by LASSO for different values of lambda. Among selected topics for all three variations of lambda, first lag of inflation expectations is selected. Column RMSE_LASSO_MIN uses the value of lambda that is equal to the minimum value of lambda chosen by cross-validation, while column RMSE_LASSO_LSE is based on the model where lambda is within one standard error. Column RMSE_LASSO_BIC is based on the lambda which is chosen using information criterion. Last two columns show the errors for benchmark AR (1) model AR (2) model, normalized relative to AR (1). Given the sparsity across normalized errors for different forecast horizons, as well as in the number of topics selected by LASSO, it can be noted that LASSO models other than that based on its minimum value are less accurate and fail to outperform the naïve models.

Table 2. RMSEs of h-period inflation expectations forecasts using different values of lambda in LASSO model, as well as AR (1) and AR (2) models. All values are normalized relative to AR (1) benchmark.

h	n_min	n_lse	n_bic	RMSE_LASSO_MIN	RMSE_LASSO_LSE	RMSE_LASSO_BIC	RMSE_AR1	RMSE_AR2
1	5	2	4	0.6	3.4	0.4	1	6
2	5	2	4	0.7	1.8	0.7	1	1.9
3	6	2	48	0.9	3.5	5.9	1	1.8
4	5	2	50	0.8	3.7	6.1	1	1.8
5	5	2	4	0.8	4.5	0.7	1	1.8
6	5	2	44	1	4.6	7.1	1	1.6
7	5	2	47	1.1	5.1	7.4	1	1.7
8	5	2	41	1	5	7.5	1	2.1
9	5	2	41	1.2	4.9	6.8	1	2.9
10	4	3	2	1.6	1.3	1.5	1	1.9
11	3	2	2	1.3	1.3	1.5	1	1.8
12	3	2	2	1.5	1.5	1.6	1	1.8

The model obtained from RMSE_LASSO_LSE includes less topics but shows poor forecasting performance. Similarly, the model from RMSE_LASSO_BIC includes even more predictors, particularly in the intermediate horizons, however, shows even worse performance. In the shorter forecasting horizons, the number of chosen topics is four, which is closer to five from the minimum lambda model, and the forecast accuracy improves. These analyses demonstrate that the optimal number of topics to predict inflation expectations up to five quarters ahead are between four and five. This also suggests that the LASSO regression, using minimum lambda as the best lambda, results to simpler model without compromising much the model performance on the test data.

It is also of interest to look how the same news data and model can be used to predict the headline inflation. We computed forecast errors for different horizons and models compared to benchmark AR (1) for annual rate of inflation and its quarterly rate. Results, not included in this chapter, but available from authors upon request suggest that while the LASSO model built using pre-selected news topics does not outperform the benchmark models, it can however be used as a forecasting model with similar forecast accuracy as those naïve models. This means that the model obtained with LASSO regression does at least as good a job fitting the information in the data as the more complicated one.

References

Alan S. Blinder & Alan B. Krueger . What Does the Public Know about Eco-nomic Policy, and How Does It Know It? *Brookings Papers on Economic Activity, Economic Studies Program, The Brookings Institution* 35(1), 327-397

Francesco D'Amuri & Juri Marcucci (2017): The predictive power of Google searches in forecasting US unemployment*, International Journal of Forecasting*, Vol. 33, No. 4, pages 801-816.

David M. Blei, Andrew Y. Ng & Michael I. Jordan (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3, 993–1022.

Erik Cambria & Bebo White: Jumping NLP Curves (2014). *A Review of Natural Language Processing Research, proceedings of Research Review Article IEEE Computational intelligence magazine*, 9, pp. 48.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin.

Kristin J. Forbes, Lewis Kirkham & Konstantinos Theodoridis (2017). A Trendy Ap-proach to UK Inflation Dynamics. *Bank of England Working Paper* 49.

Neil Gerstein, Bart Hobijn, Fernanda Nechio & Adam H. Shapiro (2019). The Brexit price Strike. *FRBSF Economic Letter*, Federal Bank of San Francisco.

Leif A. Thorsrud (2018). Words are the new numbers: A newsy coincident index of business cycles*, Journal of Business & Economic Statistics*.

Marcelo C. Medeiros, Gabriel F. R. Vasconcelos, Álvaro Veiga & Eduardo Zil-berman (2019). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods, *Journal of Business & Economic Statistics*.

Kristoffer P. Nimark and Stefan Pitschner (2018). News Media and Delegated Infor-mation Choice. CEPR Discussion Papers 11323, *C.E.P.R. Discussion Papers*.

Joonas Tuhkuri: Forecasting Unemployment with Google Searches (2016). *ETLA Working Papers*, No 35.

Yu L, Zhao Y, Tang L and Yang (2018). Online big data-driven oil consumption fore-casting with Google trends, *International Journal of Forecasting.*

Rickard Nyman, David Gregory, Sujit Kapadia, Paul Ormerod, David Tuckett & Robert Smith (2015). News and narratives in financial systems: exploiting big data for systemic risk assessment, mimeo.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11359

Citizens' attention in Madrid City through the study of personalized records

Pilar Rey del Castillo
Instituto de Estudios Fiscales, Ministry of Finance, Spain.

Abstract

The datification of our daily lives in the Big Data era is producing a huge amount of information about processes and activities that were previously invisible or at least difficult to grasp, leading to new opportunities and challenges for analysis.

Examples of some data available are the tens of million of Personalized Attention Records that can be downloaded from the open data portal offered by the local government of Madrid City. These records become a sort of counterpart from the call receiver's perspective of the Call Detail Records produced by telecom providers. They are stored as a result of a front office tool retaining some information from a range of different communication channels to manage the interaction with users.

The paper explores the data contained on these Personalized Attention Records to help improve customer attention services. It emphasizes the study of the topics that concern the citizens and the different channels dealing with the services, using Natural Language Processing and other tools.

Keywords: *Big Data; Call Records; Natural Language Processing.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

The Personalized Attention Records (PARs) of Linea Madrid are between the datasets made available by the local government of Madrid City in its open data portal https://datos.madrid.es/portal/site/egob. These records may be considered as a sort of counterpart from the call receiver's perspective of the Call Detail Records produced by telecom providers. The source for the PARs is the Customer Relationship Management (Buttle and Maklan, 2015), a front-end tool offering an interest oriented management solution. It gathers the data from different communications channels: the 26 citizen attention offices distributed by borough, the 010 phone number, the website chat, the Facebook account and the Twitter account @lineamadrid. The volume of the downloaded information, more than 44 million of records from 2014, cannot be processed using conventional statistical software and requires procedures specially developed for this purpose. Apache Spark (Zaharia et al., 2016), an open source analytics engine for Big Data processing has been used for the first steps of collecting and pre-processing data. Besides this, Python software (Van Rossum & Drake, 2009) and Scikit-learn (Pedregosa et al., 2011), a free software machine learning library for the Python programming language have been used for further calculations and analysis.

Each record contains a number of variables that have been changing through time. Some of these variables, such as the responsible worker or whether the issue has been addressed to another instance, are mostly interesting for administrative purposes. But there are other variables whose analysis may help to improve customer attention services. Besides the reception and register date, this paper focuses on the study of variables remaining through time, such as the topics that concern the citizens and the different channels dealing with the services.

Table 1. Examples of topic description variables in Personalized Attention Records.

Tipo 1	**Tipo 2**	**Tipo 3**	**Tipo 4**
Información general	Administración Pública	Administración estatal	
Movilidad	Madrid Central	Alta personas	
Identificación electrónica	Acceso a Carpeta Ciudadano	Alta	
Tasas e impuestos	IBI	Consulta/Información	Voluntaria
Cita Previa	Cita Previa	Asignar cita previa	
Movilidad	Multas	Pago con tarjeta	Voluntario
Padrón municipal	Justificantes empadronamiento	Volante empadronamiento	
Registro	Registro	Anotación	
Avisos	Avisos	Alta/Reiteración	

There is a specific variable reflecting the channel while the topic is spread out over four variables (tipo1 to tipo4). The target of using four variables to collect the topic seems to be obtaining a hierarchical description allowing for detailed information. But in practice the data have been filled in in various ways as can be seen in Table 1.

Daily indicators of the number of requests or questions received by channel and topic will be computed to provide an idea of the manner in which citizens' attention is managed by municipality services. For this purpose, the topic description variables need to be previously treated by Natural Language Processing tools.

The remainder of this paper is organized as follows. Next section describes the first steps of processing the records, making them homogeneous and obtaining a realistic classification of topics; section 3 presents and analyses the results obtained for the period between January 2014 and March 2020; and finally, a number of remarks and conclusions are presented in Section 4.

2. Processing of the Personalized Attention Records

The datasets including the records registered in the month are made available in the Madrid City open data portal after the end of each month. These files include the information of a number of variables varying in time making around 700 000 data points for each year and each variable.

The first step consists always on detecting and correcting possible logical inconsistencies in the data. For instance, for each new dataset, changes on date formats and variables appearing or disappearing are frequently found and must be previously detected to allow for subsequent homogeneous treatments. Likewise, once within the file, common spelling errors in string variables that have a fixed number of categories can be detected and corrected.

After the previous corrections, the selected data reflect the reception date, the channel receiving the request and the topic description variables. These last four variables (tipo1 to tipo4) must then be treated to obtain practical categories for classifying the topics which the requests refer to. For this purpose, some steps of Natural Language Processing (Jurafsky & Martin, 2008) have been carried out:

- Concatenation of variables tipo1 to tipo4 into one topic string, and conversion into lower case letters.
- Tokenization of the topic string into words.
- Elimination of duplicate words.
- Elimination of Spanish stop words (most common words that are filtered out).

These steps result in a list of significant words reflecting the topic for each record. A wordcloud is shown in Figure 1, which consists of a visual representation of the words in the topic descriptions where the size of each word is proportional to its frequency (Halvey and Keane, 2007).

Figure 1. Wordcloud for the words reflecting the topic for the Personalized Attention Records.

The first attempt to obtain a classification for the topics was done using Latent Dirichlet Allocation (Blei et al., 2003), a method for unsupervised classification that try to identify the topics best describing a set of documents. It builds a probabilistic model where each topic is characterized by a distribution over words, and documents are assigned probabilities to come from each topic. In the case of the PARs, the list of words for each record is too short, producing not interesting topics.

The next step has been finding the most frequent lists of one or more words by means of the Frequent Pattern Growth Algorithm, an efficient method of finding association rules (Witten et al., 2016). A classification on 15 classes has been established using these lists: *Bicycle, Culture and Sport, Elderlies, Electronic Signing, Environment, Employment, Foreigners, Health, Central Madrid*[1]*, Mobility, Other, Penalties, Register, Taxes,* and *Urban Planning*. In this way, each record is classified in one of the categories depending on the words in its topic description.

[1] New big traffic-restricted area

3. Analysis of the results

For each record the data include now the date, the channel receiving the request and the topic it refers to. A first question to raise is about the relationship between channel and topic, whether there is any kind of association between these two nominal variables. The most popular measure of relationship between this type of variables is Cramer's V (Cramér, 1946) that takes values between 0 and 1, with values closer to 1 indicating a greater association. In this case its value is 0.20, reflecting a low-medium level of relationship.

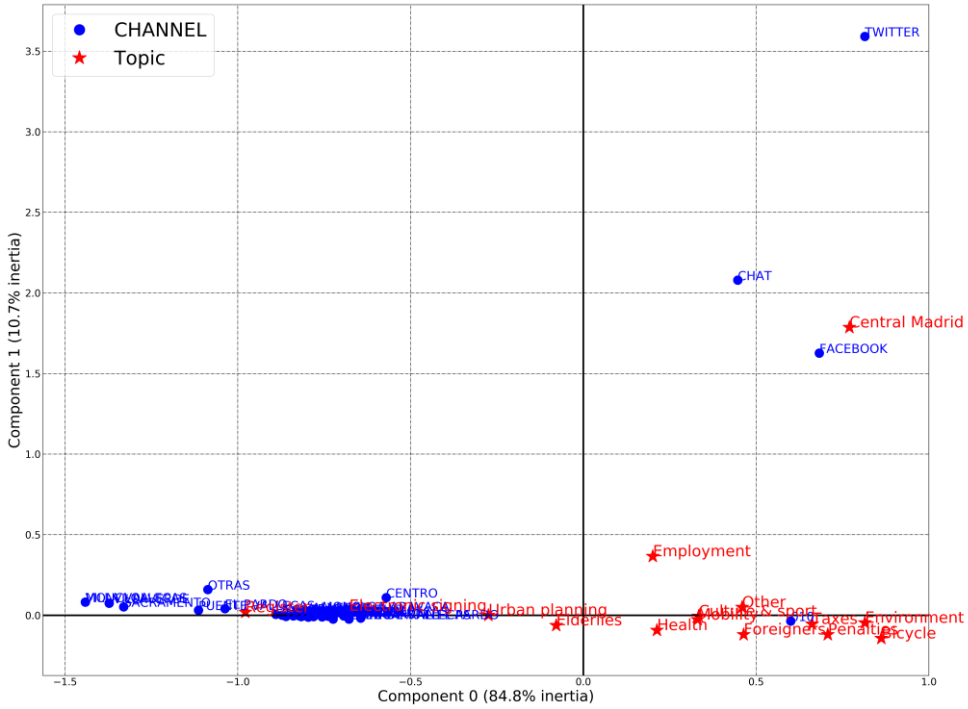


Figure 2. Principal coordinates for Channel and Topic categories.

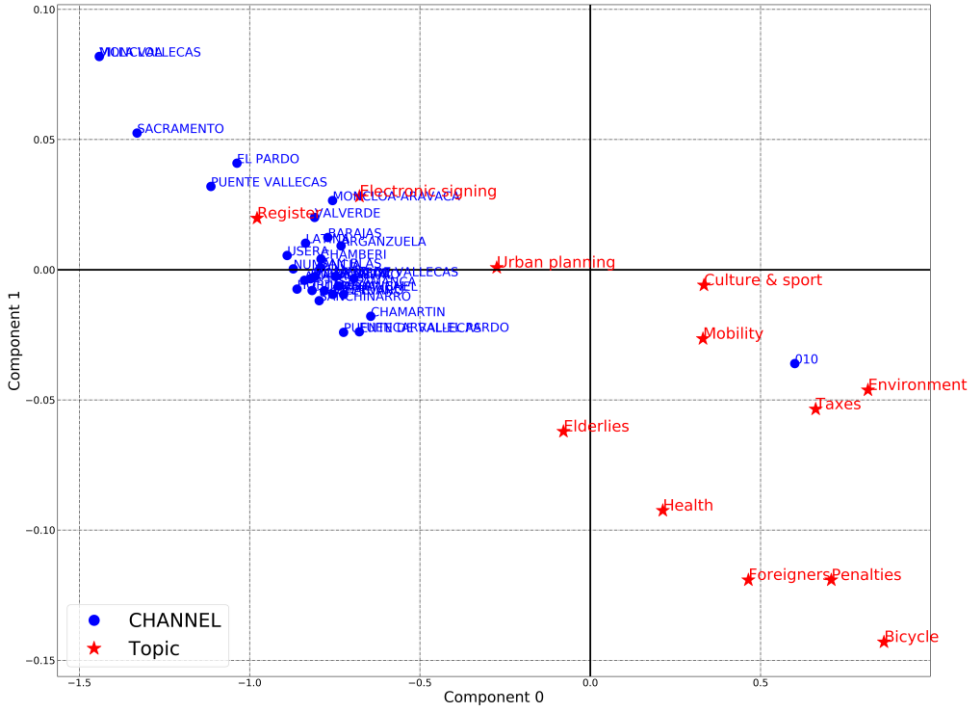


Figure 3. Principal coordinates for Channel and Topic selected categories.

Another way to study the association between channel and topic is the simple correspondence analysis (Greenacre, 2007), a graphical visualization of the rows and columns of a two-way contingency table as points in a low-dimensional space, where the positions of the row and column points are consistent with their associations in the table. The symmetrical plot including the principal coordinates for channels (rows) and topics (columns) appears in Figure 2. What can be said at first glance is that *Twitter, Chat* and *Facebook* channels have a similar profile of requests per topic, and that the requests with *Central Madrid* topic are more frequently received by these channels. Figure 3 amplifies Figure 2 to show other categories in more detail. It can be seen that the 26 borough offices have similar requests profiles and receive more frequently the requests related to *Register, Electronic signing*, and *Urban planning* while the *010* phone number receives mostly requests related to *Environment, Taxes,* and *Mobility*. The requests are unevenly distributed by channel, with the *010* phone number receiving more than 55%. The distribution per topic through all periods is shown in Figure 4. It must be taken into account that some topics such as *Central Madrid* and *Bicycle* have more recently appeared.

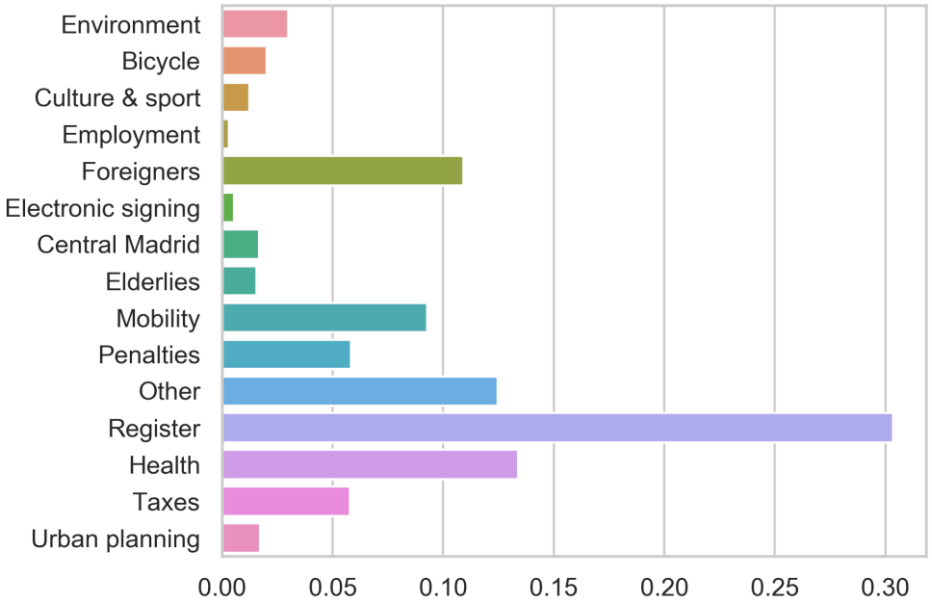


Figure 4. Proportion of requests per topic through January 2014 - March 2020.

The next step is to compute the proposed indicators. Figure 5 shows the time series of the total number of requests dealt with per day with a trend line, for the period between January 2014 and March 2020. Daily indicators for channels and topics have been similarly computed resulting in diverse rhythms, seasonal behaviors, trends, changes, and evolving behavior. The paper just presents some characteristics of the total series.

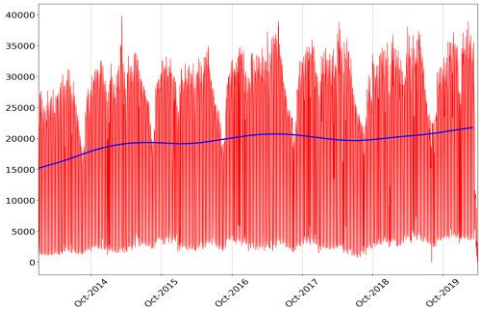


Figure 5. Total number of requests per day and trend line.

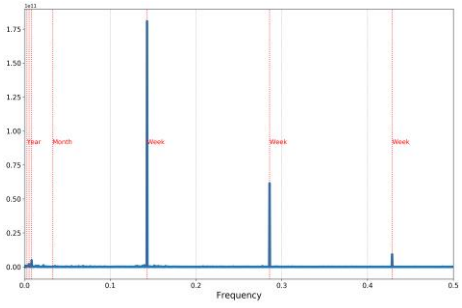


Figure 6. Periodogram spectrum estimates.

It can be seen in Figure 5 that the day-to-day movement has a lot of noise. There is also a strong pattern of seasonal decreasing in August and less markedly in December. Its periodogram spectrum estimated using Welch's method (Welch, 1967) is shown in Figure 6,

where the peaks in the spectrum indicate the frequencies of cyclical movements. The highest frequencies correspond to weekly periods, there are small frequencies for annual periods, and the frequencies are only just different from zero for monthly periods. Therefore, the most important cyclical oscillations correspond to weekly periods although these oscillations can hardly be seen in Figure 5 due to the big number of data.

To perform the seasonal adjustment of the series, a plugin of JDemetra+ version 3.0 (Eurostat, 2017) for the seasonal adjustment of daily and weekly series developed for testing purposes has been used. Boxplots of the weekly seasonality distribution are shown in Figure 7, where it can be seen that this seasonality is very stable and regular with decreasing growth the weekdays and decay on weekend. The annual seasonality average is shown in Figure 8. It results very irregular on its side because it has been calculated only with less than seven years and must be taken with caution. It reflects decays on periods matching dates of celebrations or holidays and days around them (first of May, May the 15, October the 12, first of November, …) and during the summer holidays (July - August).

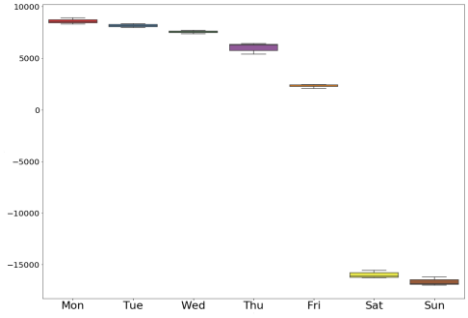


Figure 7. Weekly seasonality boxplots.

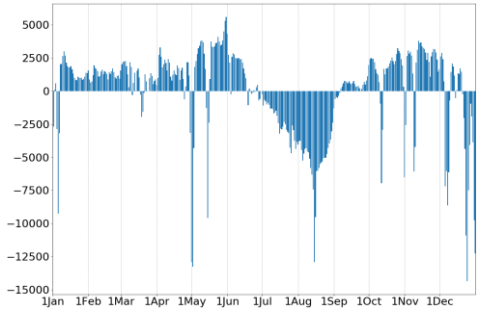


Figure 8. Annual seasonality average.

4. Last remarks

The previous sections have shown a procedure to build a classification for the topics to which the PARs refers to, and also how to compute daily indicators for the different channels and topics. The indicators can be analysed from the time series perspective to learn about the levels of workload in the channels, the seasonal behavior per topic and other issues.

There are many different ways in which this information may help to improve the attention to citizens. Although there have been shown just some examples for the global requests due to the limitation of space, other possible cases are: computing an approximate idea of the average level of occupancy of each channel during the week by calculating the average of the number of calls per day of the week and hour, and later dividing these averages by the maximum found number at this channel in an hour; executing a more precise natural language analysis of the typical issues requested by users within a topic to build a software application

for automating the customer service by conducting on-line chat conversations instead of providing direct contact with a live human agent.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3 (4–5), 993–1022.

Buttle, F., & Maklan, S. (2015). Customer Relationship Management: Concepts and Technologies. NY: Routledge, Business & Economics.

Cramér, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.

Eurostat (2017). JDemetra+ Reference Manual version 2.2. Retrieved from https://ec.europa.eu/eurostat/cros.

Greenacre, M. (2007). Correspondence Analysis in Practice (2nd ed.). London: Chapman & Hall/CRC.

Halvey, M., & Keane, M. T. (2007). An Assessment of Tag Presentation Techniques. Poster presentation at World Wide Web Conference 2007. Calgary, Canada.

Jurafsky, D. & Martin, J. H. (2008). Speech and Language Processing (2nd ed.). NJ: Pearson Prentice Hall.

Pedregosa, F., Varoquaux, G., Gramfort, A. , Michel, V., Thirion, B., Grisel, O., Blondel, M.,Prettenhofer, P., Weiss, R. , Dubourg, V. Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825—2830.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Welch, P. D. (1967). The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics* 15(2), 70–73.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Burlington, MA: Morgan Kaufmann.

Zaharia, M., Reynold, S. X., Wendell, P. Das, T., Armbrust, M., Dave, A.,Meng, X., Rosen, J.,Vernkataramen, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S. & Stoica, I. (2016). Apache Spark: A Unified Engine for Big Data Processing. *Communications of the ACM*, 56(11), 56-65.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11410

Investigating the impacts of street environment on pre-owned housing price in Shanghai using street-level images

Waishan Qiu[1], Xiaokai Huang[2], Xiaojiang Li[3], Wenjing Li[4], Ziye Zhang[5]
[1]Department of City and Regional Planning, Cornell University, USA; [2]Temple University, USA; [3]Graduate School of Design, Harvard University, USA; [4]Center for Spatial Information Science, The University of Tokyo, Japan; [5]Cornell Institute for China Economic Research, Cornell University, USA.

Abstract

Studies considering street environment quality's impact on housing value were limited to top-down variables such as the green ratio measured from satellite maps. In contrast, this study quantified street views' impacts on the value of second-hand commodity residential properties in Shanghai based on analysis of street view imagery. (1) It applied computer vision to objectively measure street features from largely accessible street view imagery. (2) Based on the classical urban design measures frameworks, it applied machine learning to evaluate human perceived street quality as street scores systematically, in contrast to the common practice of doing so in a more intuition-based fashion. (3) It further identified important indicators from both human-centered street scores as well as the more objective street feature measures with positive or adverse effects on property values based on a hedonic modeling method. The estimation suggested both street scores and features are significant and non-negligible. For the perceived street scores (from 0-10 scale), neighborhoods with a unit increase in their "enclosure" or "safety" score enjoy price premium of 0.3% to 0.6%. Meanwhile, streets with 10% greater tree canopy exposure are attributable to a 0.2% increase in the property value. This study enriched our current understanding at a micro level of the factors that impact property values from the perspective of the built environment. It introduced human-centered perception of street scores and objective measures of street features as spatial variables into the analysis of neighborhood attribute vectors.

Keywords: *Machine learning; Property value; Shanghai; Street view imagery.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

1.1. Street Quality and Property Values

The quality of streets is among those factors that have attracted attention in terms of its impact on the surrounding property values. Scholars have made many efforts to investigate the impact of streetscapes especially those urban trees on the property values. Most existing research focuses on objective metrics such as the numbers of trees, the size of tree covers, the distance to large green areas, or the total greenery measured from satellite images, failing to incorporate the importance of human perceptions. Killicoat et al. (2002) identified trees as important amenities creating an environment conducive to real estate transactions. Using a hedonic price model (HPM), Sander et al. (2010) implied that 10% more tree cover on parcels increased sales price by 0.48% in Minnesota. Besides trees, the impacts of other features on the property value, such as open space (Anderson & West, 2006), street lighting (Willis et al., 2005), and surface street traffic (Larsen & Blair, 2014) have been studied separately with the HPM approach.

1.2. Street Environment Measures

Physical features alone may not represent people's overall image of street quality, because the cumulative effect is greater than the total of each. Streetscape affecting perceived street environment quality should be defined and measured empirically (Yin, 2014). Cervero & Kockelman (1997) first proposed the framework of 3Ds- density, diversity and design, and used gross quantities such as building height, block length and street width to measure the design quality of streets. However, people do not experience a street from these gross metrics. More operational definitions and measurement protocols about perceived quality were required (Brownson et al., 2004; Ewing & Clemente, 2013a; Laaksonen et al., 2006; Pikora et al., 2003). Expert panels were recruited by Ewing et al. (2006) to rate five streetscape qualities based on video clips, namely imageability, visual enclosure, human scale, transparency and complexity. Physical features in the video clips were also quantified, and their relationships with the ratings from expert panels were statistically analyzed (Ewing et al., 2006). Conventional method such as visual collage or phone surveys to collect people's perceptions has many problems. First, the reliability of the operation is questionable (Nasar, 1990), as individual differences are one of the problems to make the evaluation impractical (Ewing & Handy, 2009). Second, the low-throughput method results in constrained application to larger geographic contexts. Limited by the high survey cost or the poor spatial resolution, a segment of the city was often utilized as a sample area, making their conclusions subject to the particular conditions in that area and weak in generalizability (Seiferling et al., 2017).

1.3. Novel Streetscape Measures

Most recently, studies in this regard started to exploit multisource open data and showed significant advantages in efficiency and accuracy over traditional methods. One example is the increasing use of the street view image (SVI). Differing from satellite imagery, SVI represents a ground-based perspective of urban streets and is available for the city wide (Seiferling et al., 2017). Compared to on-site auditing, the use of SVI was proved to be efficient for large-scale objective data collection (Griew et al., 2013; Rundle et al., 2011). Meanwhile, with the advance of artificial intelligence (AI), computer vision (CV) and machine learning (ML) algorithms were developed to predict human perception of the street environment from images. Naik et al. (2014) predicted the perceived safety score using labelled SVI as the training data from an online survey with more than 7000 participants. Chen et al. (2020) compared the non-linear relationship between housing prices and GVI in Shanghai and found higher GVI were significantly associated with more expensive houses.

Despite these meaningful explorations, our understanding of the impact of the street environment on economic outcomes is still limited by the drawbacks in the literature such as small samples, inconsistent ratings, and lack of human perception measures. This study aimed to investigate the impact of perceived streetscape on property values for the city-wide Shanghai, using a reliable and high-throughput approach with the advance of AI. It achieved at least four key contributions. First, it evaluated the feasibility of utilizing SVI and CV for auditing urban environments at a large scale, quantifying percentile index or counts for 30 street features with 25,276 SVIs. Second, it developed an online survey collecting the collaborative perceptions of street quality with 100 participants. Third, it trained efficient ML models to evaluate three important perceived street qualities, namely visual enclosure, complexity and aesthetics. Fourth, it identified the effects of perceived qualities and objective percentile index of streetscape with positive or adverse effects on property values using more than 40,000 housing transactions. By providing a more comprehensive understanding of the relationship between perceived streetscape qualities and property values in a city-wide scale, this study would inform policy-making and urban design.

2. Data and Methods

2.1. Hedonic Price Model

The hedonic price model (HPM) was widely adopted to investigate the impacts of built environment features on housing prices (Chen et al., 2020; Larsen & Blair, 2014; Zhang & Dong, 2018). It assumes that housing is a heterogeneous good, and the determinants of its price can be investigated by regressing the housing price on three types of independent variables capturing the property's structural, locational and neighborhood attributes (Rosen,

1974). Our focus is on a subgroup of attributes relevant to the street quality, and thus we expand the conventional HPM as follows:

$$\ln\text{PRICE} = \alpha + \beta_1 \mathbf{STRU} + \beta_2 \mathbf{LOCA} + \beta_3 \mathbf{NEIG} + \beta_4 \mathbf{STRE} + \varepsilon, \quad (1)$$

where lnPRICE was the natural logarithm of the transaction price; α is the regression constant; β_1 to β_4 are the estimated coefficients of the structural **(STRU)**, locational **(LOC)**, neighborhood **(NEIG)**, and streetscape (**(STRE)**) attributes respectively; ε is the error term.

2.2. Data Collection and Processing

2.2.1. Housing Transactional Price

This study collected 65,000 transaction records of preowned apartments occurring in 2019 within Shanghai from Lianjia.com. Besides the sales price, the data also included housing structure attributes and property's geolocation. The dataset was cleaned for (1) records whose price did not make sense, e.g. being zero or much higher or lower than the neighbors' mean; and (2) records whose property attributes were missing. 40,159 geo-tagged records were kept, with an average sales price at about 57,000 RMB/m^2.

2.2.2. Structural Attributes

For structural attributes, selected continuous variables included floor area, numbers of bathroom, total numbers of floor, the year of construction. Categorical variables of relative floor level in the building, unit layout, building type, facing direction, structure type, interior decoration, availability of elevators were transformed to dummies.

2.2.3. Locational Attributes

Notably, a large body of literature have identified the pattern when distances to the city center increase, housing prices decrease (Bourassa et al., 2010; Chen et al., 2020; Rosen, 1974). We calculated the Euclidean distance from each property (1) to the Bund - CBD of Shanghai and (2) to their nearest county center as locational attributes.

2.2.4. Neighborhood Attributes

The neighborhood variables measured the (1) density, (2) distance to and (3) accessibility to urban services. The living amenities density was the number of services such as retailing, restaurants, cafes, groceries, hospitals and gyms per km^2 within the neighborhood's administration boundary. Distances variables calculated the Euclidean distance to the closest metro station and school. Accessibility to metro stations and schools was also adopted and measured by the numbers of schools and metro stations can be reached to within 5 kilometer and 1 kilometers respectively.

2.2.5. Street environment features

Motivated by the classical measurement protocols of Ewing et al. (2006), this study chose "visual enclosure" and "complexity", and modified the term "imageability" to "aesthetics" to present the perceived streetscape qualities in three different dimensions. Specifically, (1) visual enclosure refers to the extent streets are visually defined by features such as trees, walls and buildings; (2) complexity refers to the visual richness of a place, depending on the variety of and numbers of elements such as human activity, signage, street furniture, greenery, and buildings; (3) borrowing the term "imageability" which was a categorical variable, instead we used "aesthetics" to present the overall visual quality of a street being comfortable and distinct (Ewing et al., 2006). Meanwhile, many literatures identified urban trees as a significant factor affecting property value by using the percentile index GVI. We selected GVI as the objective measure, whose value was given by the ratio of pixels associated with tree canopy to the total pixels in an image:

$$\mathrm{GVI} = \mathrm{Pixels}_{\mathrm{greenery}} / \mathrm{Pixels}_{\mathrm{total}}, \quad (2)$$

The process of calculating these four variables from images involved five steps: (1) sampled SVI sites, (2) downloaded SVI, (3) collected people's collaborative ratings as training data's Y labels, (4) extracted pixel ratios of different streetscape features as training data's X variables, and (5) trained ML models and predicted perceived scores. First, given that many transactional data had the duplicated coordinates, its SVI data were sampled at 50-meter intervals along the centerline of all public streets within the property's 1-km radius. The sampling was processed in ArcGIS. In total, we sampled 30,000 SVI sites. Second, for each SVI sample site, its SVI was acquired by inputting its coordinates into Baidu Street View API. We controlled consistent camera settings, same size (600 by 300 pixels) and similar shooting time (summer and fall 2017) for each SVI to ensure the consistent perspective and to eliminate the seasonal variance in street environments (**Fig.1b**). Due to the limited data availability, not all sampled sites have SVI, as a result we downloaded 25,276 valid SVI. The values for the missing sample sites would be estimated through spatial interpolation detailed in the next section.

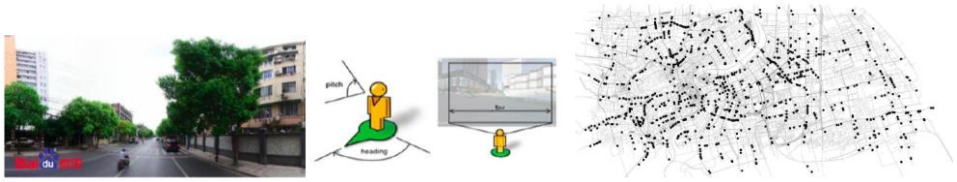


Figure 1. Downloading Baidu Street View Image.
(a) A typical SVI downloaded for this study, whose "point of view" was located in the middle of the road, being paralleled to the road's direction. (b) The camera settings were controlled by "heading", "FOV", "pitch" and "resolution". In this study, the camera heading was set the same as the sample site's street center line tangent degree; the FOV (horizontal field of view) was set 120 degree, the pitch (specifies the up or down angle of the camera relative to the Street View vehicle) was 0 degree, and the resolution was 600 x 300 pixels, for all images. (c) Training images were sampled across a wide range of geographical locations in Shanghai.

Third, to acquire training data of people's preferences on streetscapes from SVI, we adopted the high-throughput method developed from emerging studies (Naik et al., 2014; Salesses et al., 2013). Specifically, we developed an online survey platform on which participants were shown two randomly chosen SVIs in Shanghai. Participants were asked to click on the SVI they preferred in response to three evaluative questions in regard to perceived street qualities, namely the Visual Enclosure, the Complexity and the Aesthetics. For example, for Visual Enclosure, we gave a brief definition on Visual Enclosure. Then participants were asked the question "Which place do you feel having a higher level of visual enclosure". Noted that, to ensure the training images would cover a wide range of streetscapes from city center, suburban to countryside, we randomly sampled 500 images across the whole region of Shanghai **(Fig. 1c)**. In total 100 unique participants from Shanghai ranked the images using 8426 pairwise clicks. These preferences were converted to ranked scores for each image with Microsoft Trueskill algorithm (Naik et al., 2014). On average each unique image was compared to 16.8 other images, which was sufficient to lead Trueskill converge to stable ranked scores (Naik et al., 2014). The ranked scores were then normalized to a 0-10 scale. These 500 images with three ranked scores as labels, became the training data from which ML algorithms were trained to predict the collaborative perceptions of the streetscape.

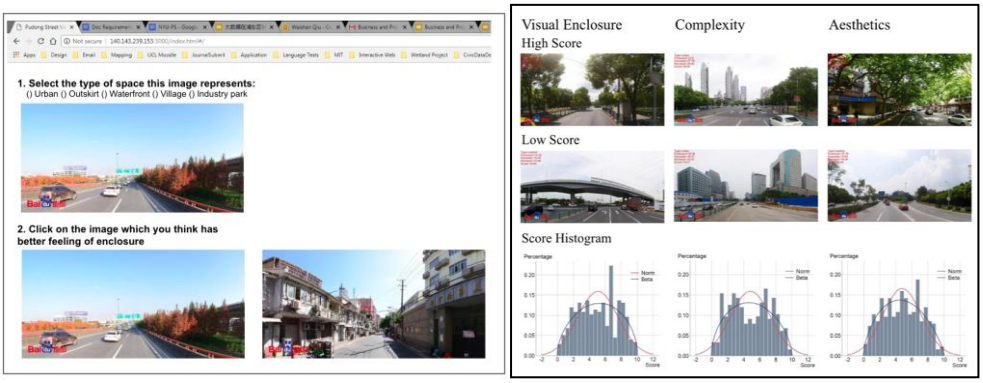


Figure 2. Collaborative Image of Streetscape with an Online Survey Platform (a) The survey system asked participants to choose from the pairwise SVIs in response to questions. (b) High score, low score example images, and the histogram of score distribution, for each of the three perceived street qualities. From the high and low score examples, people seemed to prefer streetscapes with less sky exposure and more greenery.

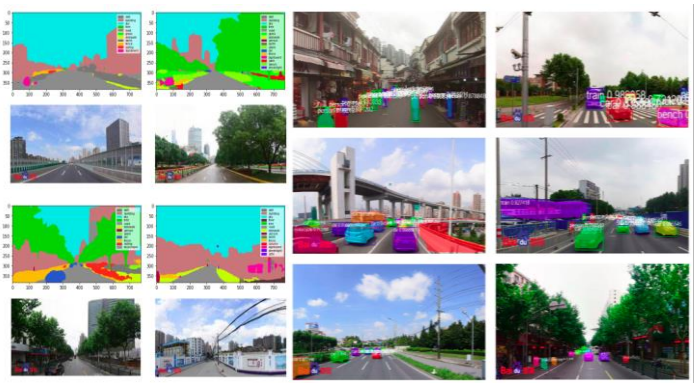


Figure 3. CV Segmentation Results (a) Pairwise PSPNet semantic segmentation results and its input SVI (b) Mask R-CNN instance segmentation results counting numbers of streetscape objects appearing in the SVI

Fourth, we quantified the pixel ratios or amounts of 30 street features from the 500 training images using pre-trained CV models. These features, including the building, sky, tree, curbs, roads, grass, glass and others, were conceived significant effects on human perceived street qualities by Ewing et al. (2006). On one hand, we applied Pyramid Scene Parsing Network (PSPNet) to calculate the pixel ratios of 30 kinds of features (Zhao et al., 2016). On the other hand, for five of the features, namely the car, people, bike, motorbike and chairs, an indicator of pixel ratio makes less sense than the absolute counts. So we applied pre-trained Mask R-CNN to count their amounts (He et al., 2017). These 30 features became the explanatory variables to predict three perceived scores in the next step.

Fifth, we trained ML models including K-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), decision tree, and gradient boost to predict the three perceived scores from the 30 features extracted from SVI. The 500 labelled images from the online survey were split to 400 and 100 for training and testing purposes. To choose the optimal model, we compared model performances in regard to the R-square , the root mean square error (RMSE) , and the mean absolute error (MAE). SVM outperformed other models in predicting enclosure and complexity score, while RF performed best in predicting aesthetics score. Regarding the accuracy, SVM predicted enclosure and complexity scores with an MAE of 1.51 and 1.35 respectively, while RF predicted aesthetics score with a 1.19 MAE (**Table 3**). The accuracy of ML model was considered acceptable, limiting by the small training dataset, the error magnitude of 1.2 to 1.51 would not offset predicted scores far from true value with regard to the 0-10 scale scoring system. The global importance of individual streetscape features were compared using Python Scikit-learn's Tree Based Regressor. The algorithm computes how much each variable contributes to decreasing the weighted impurity during the training, thus providing the importance score (Chen et al., 2020; Naik et al., 2014). Pixels of trees ranked second, justified that GVI was an appropriate variable to represent the objective measure of streetscape.

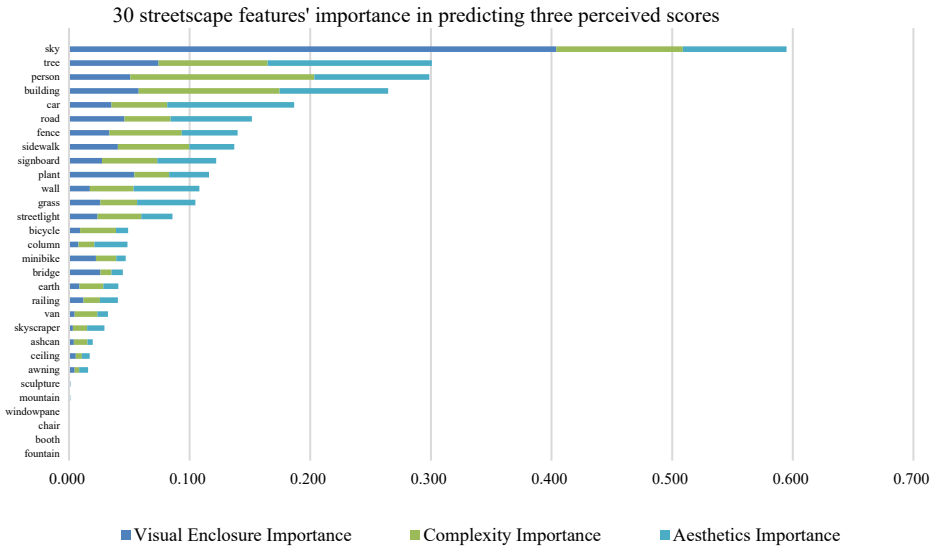


Figure 4. Feature importance in predicting perceived scores.

Reference

Anderson, S. T., & West, S. E. (2006). Open space, residential property values, and spatial context. *Regional Science and Urban Economics*, *36*(6), 773–789. https://doi.org/10.1016/j.regsciurbeco.2006.03.007

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods. *Journal of Real Estate Research*, *32*(2), 139–160.

Brownson, R. C., Chang, J. J., Eyler, A. A., Ainsworth, B. E., Kirtland, K. A., Saelens, B. E., & Sallis, J. F. (2004). Measuring the Environment for Friendliness Toward Physical Activity: A Comparison of the Reliability of 3 Questionnaires. *American Journal of Public Health*, *94*(3), 473–483. https://doi.org/10.2105/AJPH.94.3.473

Brownson, R. C., Hoehner, C. M., Day, K., Forsyth, A., & Sallis, J. F. (2009). Measuring the built environment for physical activity: state of the science. *American Journal of Preventive Medicine*, *36*(4 Suppl), S99–S123.e12. https://doi.org/10.1016/j.amepre.2009.01.005

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, *2*(3), 199–219. https://doi.org/10.1016/S1361-9209(97)00009-6

Chen, L., Yao, X., Liu, Y., Zhu, Y., Chen, W., Zhao, X., & Chi, T. (2020). Measuring Impacts of Urban Environmental Elements on Housing Prices Based on Multisource Data—A Case Study of Shanghai, China. *ISPRS International Journal of Geo-Information*, *9*(2), 106. https://doi.org/10.3390/ijgi9020106

Donovan, G. H., & Butry, D. T. (2010). Trees in the city: Valuing street trees in Portland, Oregon. *Landscape and Urban Planning*, *94*(2), 77–83. https://doi.org/10.1016/j.landurbplan.2009.07.019

Dubé, J., & Legros, D. (2014). Spatial econometrics and the hedonic pricing model: what about the temporal dimension? *Journal of Property Research*, *31*(4), 333–359. https://doi.org/10.1080/09599916.2014.913655

Ewing, R., & Clemente, O. (2013). *Measuring Urban Design: Metrics for Livable Places*. Island Press/Center for Resource Economics. https://www.springer.com/gp/book/9781610912099

Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design*, *14*(1), 65–84.

Ewing, R., Handy, S., Brownson, R. C., Clemente, O., & Winston, E. (2006). Identifying and Measuring Urban Design Qualities Related to Walkability. *Journal of Physical Activity & Health*, *3*(s1), S223–S240. https://doi.org/10.1123/jpah.3.s1.s223

Griew, P., Hillsdon, M., Foster, C., Coombes, E., Jones, A., & Wilkinson, P. (2013). Developing and testing a street audit tool using Google Street View to measure environmental supportiveness for physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, *10*(1), 103. https://doi.org/10.1186/1479-5868-10-103

Hara, K., Sun, J., Moore, R., Jacobs, D., & Froehlich, J. (2014). Tohme: detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning.

Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology - UIST '14, 189–204. https://doi.org/10.1145/2642918.2647403

Kelly, C. M., Wilson, J. S., Baker, E. A., Miller, D. K., & Schootman, M. (2012). Using Google Street View to audit the built environment: inter-rater reliability results. *Annals of Behavioral Medicine*, *45*(suppl_1), S108–S112.

Killicoat, P., Puzio, E., & Stringer, R. (2002). The economic value of trees in urban areas: estimating the benefits of Adelaide's street trees. *Proceedings Treenet Symposium*, *94*, 106.

Laaksonen, P., Laaksonen, M., Borisov, P., & Halkoaho, J. (2006). Measuring image of a city: A qualitative approach with case example. *Place Branding*, *2*(3), 210–219. https://doi.org/10.1057/palgrave.pb.5990058

Larsen, J. E., & Blair, J. P. (2014). Price effects of surface street traffic on residential property. *International Journal of Housing Markets and Analysis*. https://doi.org/10.1108/IJHMA-12-2012-0062

Lin, L., & Moudon, A. V. (2010). Objective versus subjective measures of the built environment, which are most effective in capturing associations with walking? *Health & Place*, *16*(2), 339–348. https://doi.org/10.1016/j.healthplace.2009.11.002

Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening*, *14*(3), 675–685. https://doi.org/10.1016/j.ufug.2015.06.006

Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore -- Predicting the Perceived Safety of One Million Streetscapes. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 793–799. https://doi.org/10.1109/CVPRW.2014.121

Nasar, J. L. (1990). The Evaluative Image of the City. *Journal of the American Planning Association*, *56*(1), 41–53. https://doi.org/10.1080/01944369008975742

Orland, B., Vining, J., & Ebreo, A. (1992). The Effect of Street Trees on Perceived Values of Residential Property. *Environment and Behavior*, *24*(3), 298–325. https://doi.org/10.1177/0013916592243002

Pandit, R., Polyakov, M., & Sadler, R. (2014). Valuing public and private urban tree canopy cover. *Australian Journal of Agricultural and Resource Economics*, *58*(3), 453–470. https://doi.org/10.1111/1467-8489.12037

Pikora, T., Giles-Corti, B., Bull, F., Jamrozik, K., & Donovan, R. (2003). Developing a framework for assessment of the environmental determinants of walking and cycling. *Social Science & Medicine*, *56*(8), 1693–1703. https://doi.org/10.1016/S0277-9536(02)00163-6

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34–55. JSTOR.

Rundle, A. G., Bader, M. D. M., Richards, C. A., Neckerman, K. M., & Teitler, J. O. (2011). Using Google Street View to Audit Neighborhood Environments. *American Journal of Preventive Medicine*, *40*(1), 94–100. https://doi.org/10.1016/j.amepre.2010.09.034

Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The Collaborative Image of The City: Mapping the Inequality of Urban Perception. *PLOS ONE*, *8*(7), e68400. https://doi.org/10.1371/journal.pone.0068400

Sander, H., Polasky, S., & Haight, R. G. (2010). The value of urban tree cover: A hedonic property price model in Ramsey and Dakota Counties, Minnesota, USA. *Ecological Economics*, *69*(8), 1646–1656. https://doi.org/10.1016/j.ecolecon.2010.03.011

Seiferling, I., Naik, N., Ratti, C., & Proulx, R. (2017). Green streets- Quantifying and mapping urban trees with street-level imagery and computer vision. *Landscape and Urban Planning*, *165*, 93–101.

Willis, K. G., Powe, N. A., & Garrod, G. D. (2005). Estimating the Value of Improved Street Lighting: A Factor Analytical Discrete Choice Approach. *Urban Studies*, *42*(12), 2289–2303. https://doi.org/10.1080/00420980500332106

Yin, L. (2014). Book Review: Measuring Urban Design: Metrics for Livable Places. *Journal of Planning Literature*, *29*(3), 273–274. https://doi.org/10.1177/0885412214531541

Yin, L., & Wang, Z. (2016). Measuring visual enclosure for street walkability: Using machine learning algorithms and Google Street View imagery. *Applied Geography*, *76*, 147–153. https://doi.org/10.1016/j.apgeog.2016.09.024

Zhang, Y., & Dong, R. (2018). Impacts of Street-Visible Greenery on Housing Prices: Evidence from a Hedonic Price Model and a Massive Street View Image Dataset in Beijing. *ISPRS International Journal of Geo-Information*, *7*(3), 104. https://doi.org/10.3390/ijgi7030104

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11413

eWOM in reward-based crowdfunding platforms: a behavioral approach

Irene Comeig Ramírez[1], Pau Sendra Pons[2]
[1]Department of Corporate Finance and ERI-CES, University of Valencia, Spain, [2]Department of Corporate Finance, University of Valencia, Spain.

Abstract

Electronic word of mouth (eWOM) plays a crucial role in influencing purchasing decisions of consumers in situations governed by asymmetric information. In this context, investors in reward-based crowdfunding platforms might modify their purchasing intentions according to recommendations of peers and/or experts. The goal of this paper is to analyze the power of eWOM to shape consumers' purchasing decisions. We do so by conducting an experiment through Amazon Mechanical Turk (AMT). This online experimental tool allows for an instant access to a large and culturally diverse subject pool, facilitating behavioral research requiring large amounts of subjects. By recreating a reward-based crowdfunding webpage and tracking how consumers' choices vary due to recommendations of other buyers and experts, this research confirms eWOM power in modifying purchasing decisions, as well as the prevalence of other buyers' recommendations over those of experts. Additionally, it is tested AMT as a crowdsourcing platform that enables scholars to carry out online research related to economics and social sciences.

Keywords: *eWOM; Internet; experimental economics; crowdfunding; crowdsourcing.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

Word of mouth (WOM) has been analyzed as an expression of interpersonal communication about products and services. Its power to influence consumer product judgment has been approached both theoretically and empirically, in particular, with the uptake of the Internet, which has broadened the existing channels of communications (Lee & Youn, 2009). It has led to the emergence of electronic word of mouth (eWOM), considered an influential instrument in the field of marketing. Ultimately, in a process of product choice, consumers search for information posted by previous customers and experts to reaffirm their original buying decisions (Erkan & Evans, 2016).

Crowdfunding has revolutionized capital raising as an alternative way of finance that connects those seeking funding for their business endeavors and philanthropic causes with individuals interested in investing or donating. Its revolutionary character allows for bypassing the intermediaries of a traditional supply chain making the funding process more transparent and democratic. Thus, it has the potential to foster innovation as it makes it easier for risky and innovative start-ups to obtain funds. Furthermore, the crowd provides feedback to the entrepreneur while interacting in the funding process, for example, delivering additional information on the actual demand for a product or about consumer preferences (Schwienbacher, 2018).

However, there exists asymmetric information between fund seekers and capital providers as the former have superior knowledge of their projects whereas the latter receive limited information (Agrawal *et al*., 2014). Due to this, electronic word of mouth (eWOM) plays a crucial role in prompting the investment decision. Often, backers are uncertain about the ability of the campaign promoters to collect enough contributions to reach the funding goal. In this context, we designed an economic experiment launched through Amazon Mechanical Turk (AMT) with the goal of analyzing the effect of eWOM on the investment decision in reward-based crowdfunding environments. Simultaneously, we tested AMT as an experimental tool for recruiting large number of subjects through the Internet. Results confirm the power of eWOM to modify purchasing decisions as well as prevalence of other buyers' recommendations over those of experts.

2. Literature Review

As explained by Belleflamme *et al*. (2014), the concept of crowdfunding consists in several individuals, reached mostly through the Internet, providing financial resources to support the success of all kinds of initiatives. It is derived from a broader concept, crowdsourcing, which encompasses outsourcing a task previously performed by an employee to a large mass of people in the form of an online open call. Specifically, the reward-based

crowdfunding model is mainly used by entrepreneurs to finance the manufacturing of new products. Rather than borrowing money from banks, funds are collected from the crowd.

Backers are always compensated either with a tangible reward –e.g. a sample of the final product– or an intangible one –e.g. having their name written in the product packaging. Ultimately, crowdfunding can play a substantial role in facilitating the flow of funds to risky and innovative start-ups as well as small and medium-sized companies, which might face serious challenges to get funded after the recent financial crisis (Cosh *et al.*, 2009). Crowdfunding platforms, as many other investment environments, are dominated by asymmetric information between fund seekers and capital providers. In this situation, it is likely that eWOM –i.e. sharing experiences and opinions through the internet– triggers herding behavior. Such conduct can be generally defined as a form of social behavior convergence aligning individual thoughts or behaviors with those of the group through non-coordinated interaction (Raafat *et al*., 2009). More specifically, herding can be rational, when observational learners make unbiased inferences from the behavior of others (Simonsohn & Ariely, 2008), or irrational, as a mere imitation process where investors just go along with the crowd. Therefore, hypotheses are formulated as follows:

H1: eWOM influences investors' beliefs and modifies initial investment decisions.

H2: Recommendations by peers will influence investors' decisions to a large extent than those of experts.

Previous empirical research has analyzed how recommendations made by peers and experts affect online product choice (Huang & Chen, 2006) as well as the impact of electronic word of mouth (eWOM) over the investment decision (Bi *et al*., 2017). However, to the author's knowledge, this is the first experiment that recreates a crowdfunding webpage with an online economic experiment in which subjects are rewarded according to performance.

3. Research Methodology

3.1. Experimental design

In order to explore how the financing decision of investors varies due to comments of peers and experts in reward-based crowdfunding platforms, we designed an economic experiment conducted through Amazon Mechanical Turk (AMT) where 847 subjects participated, 500 from the US and 347 from India. It recreated a crowdfunding webpage, such as *Go Fund Me* or *Kickstarter*, where subjects were asked to virtually contribute $15 to one of two projects aiming to publish a book. Both projects had the same budget requirement, $5,250, and deadline date, January 31st, 2019, which were kept constant. A project would be successful if it reached the budget threshold by the specified closing date. The experiment had two treatments, one *without information* and the other *with information*. In both

treatments two projects that intended to publish cookery books where shown for the experimental subjects to choose.

The experimental design was conceived to analyze how the reviews of peers and experts influenced the financing decision. Peers are considered those of equal standing to the normal public of a reward-based crowdfunding webpage, that is, investors that provide funds to projects in the expectation of receiving the promised products if the fundraiser succeeds. Conversely, experts are individuals with greater knowledge and experience about the specific product offered, in this case, cookery books. The treatment *without information* asked participants to virtually contribute $15 to one of the two projects, *Book A* or *Book B*, according to their cover. After, the treatment *with information* showed investors three reviews for each book. *Book A* had two negative comments of peers and one positive of an expert. Oppositely, *Book B* had the positive judgment of two peers and the negative one of an expert. Besides, it was indicated that both projects had raised $450 from 30 backers. Contributions of early investors were kept identical for both projects, being the reviews the only differentiating information.

3.2. Amazon Mechanical Turk (AMT) as a tool for conducting experiments

Amazon Mechanical Turk (AMT) is a crowdsourcing marketplace allowing employers –called *requesters*– to post tasks –called *HITs*, i.e. Human Intelligence Tasks– to be made by the platform's online labor market –composed of the so called *workers*– in return for a wage –called *reward*–. Requesters design the HITs either by using the templates offered or by creating their own template with Hyper Text Markup Language (HTML). Before posting a task, requesters decide the number of respondents needed, the time allotted for each respondent to complete the HIT as well as its expiration. Some of the criteria for selecting subjects within the platform can be chosen at no cost –e.g. HIT Approval Rate (%), location and number of HITs approved– whereas the rest involve additional costs –e.g. gender, vacation frequency, job function, primary news source or daily internet usage. Requesters who need respondents of a specific profile use the latter.

Previous researchers have highlighted the strengths and pitfalls associated with conducting experimental research on AMT. Paolacci *et al*. (2010) pointed out how its supportive infrastructure allows integrating various stages of the research process in a single platform. They also noted that subjects are identifiable by a *Worker ID*, what allows researchers to perform longitudinal studies. Besides that, Mason & Siddharth (2011) emphasized the access to a large and culturally diverse subject pool as a core strength that facilitates cross cultural comparisons. Interestingly, they also mentioned how AMT maximizes research efficiency while speeding up the experimental cycle. Furthermore, its low cost and built-in payment mechanism makes it competitive when compared to costly laboratory-based experiments. Despite quality concerns, Buhrmester *et al*. (2011) concluded that workers are willing to complete simple tasks in exchange for small compensation, what suggests they

are not primarily driven by financial incentives. Although they claimed there exists sensitivity between participation rates, compensation amounts and time commitments, they maintained participants could be recruited both rapidly and inexpensively.

Concerning pitfalls, Johnson & Borden (2012) noted the difficulty to ensure subjects devote a reasonable amount of time to complete an experiment. To tackle this problem, they proposed setting a response time benchmark, introducing Instructional Manipulation Checks (IMCs), firstly introduced by Oppenheimer *et al*. (2009), and choosing workers with an approval rating of 95% or above in order to minimize the likelihood of randomly selected answers. Also, it can be underlined the technical complexity associated with conducting experiments that require real-time interaction among participants. Regarding distractions faced by subjects, a survey by Chandler *et al.* (2013) pointed at interaction with other people as the main one. Data reliability is a major concern for many researchers using AMT. Several studies have tried to assess the validity of AMT for conducting behavioral research –e.g. see Buhrmester *et al*. (2011) or Holden *et al*. (2013). Overall, they tend to conclude that benefits in terms of science democratization and instant access to a large subject pool outweigh the weaknesses of an uncontrolled online experimental environment.

4. Results

As shown by Figure 1, recommendations of peers were more significant than those of experts at influencing investors' beliefs. In *Treatment I*, 59.4% of subjects chose Book A while 40.6% of them chose Book B. In *Treatment II*, Book A was only supported by 39.3% of subjects while Book B, 60.7% of them. Information released showed that Book A had two negative recommendations of two peers and a positive of an expert and the reverse was shown for Book B. Therefore, Book A was endorsed by experts and Book B by peers. The shift of subjects from financing Book A to Book B may have occurred either as a result of a rational reasoning, due to peers' recommendations being understood as a sign of quality by potential investors, or an irrational one, as a mere imitation of the crowd. In any case, it is confirmed the power of eWOM to influence beliefs and, specifically, peers prevalence over experts at modifying investors' choice.

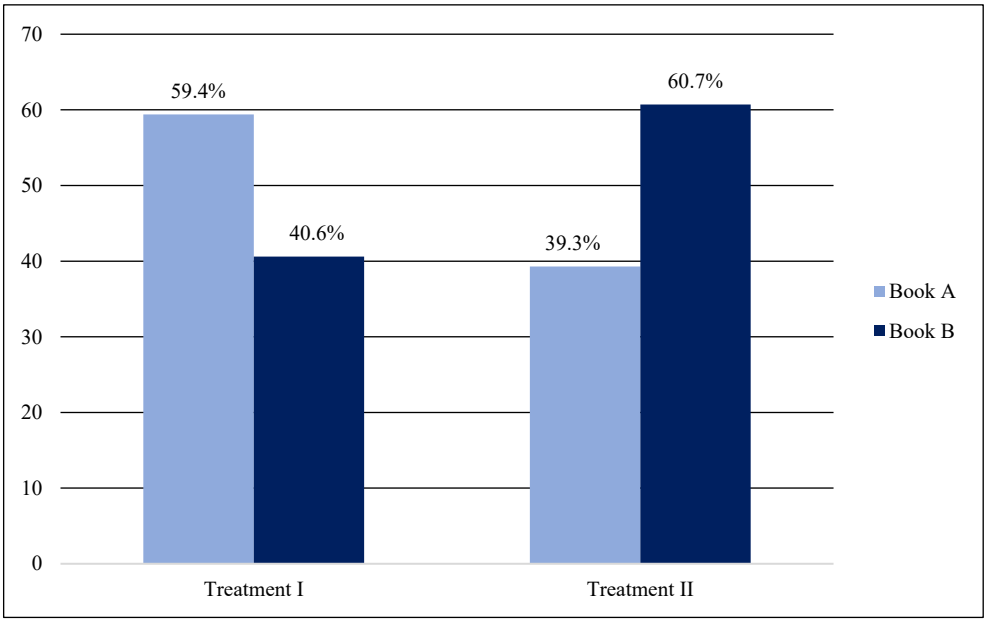


Figure 1. Testing the influence of peer and expert opinion on the investment decision.

Table 1 shows the frequencies and p-values by gender and country. On one side, Panel A displays the change in subject choice between *Treatment I* (without information) and *Treatment II* (with added information) testing the null hypothesis H_0: A/B = B/A. A/B denotes subjects changing from funding Book A in *Treatment I* to funding Book B in *Treatment II*, and B/A denotes subjects changing from funding Book B in *Treatment I* to funding Book A in *Treatment II*. On the other hand, Panel B exhibits subject choice in *Treatment II* testing the null hypothesis H_0: A = B. Regarding Panel A, all the shifts from the former investment decision to the one made after the information was released were significant. In overall terms, 212 subjects shifted from Book A to Book B while only 42 did so in the opposite direction.

Table 1. Testing the influence of peer and expert opinion on the investment decision. Frequencies and p-values by gender and country.

Panel A. Change in subject choice between Treatment I and II (with added information)

H0: A/B = A/B		**Men**		**Women**		**Men + Women**	
Country		**A/B***	**B/A**	**A/B**	**B/A**	**A/B**	**B/A**
USA	Number	52	3	83	4	135	7
	%	94.55	5.45	95.40	4.60	95.07	4.93
	Proportion test	$p < 0.0001$		$p = 0.017$		$p < 0.0001$	
India	Number	49	26	28	9	77	35
	%	65.33	34.67	75.68	24.32	68.75	31.25
	Proportion test	$p = 0.011$		$p = 0.005$		$p < 0.0001$	
USA + India	Number	101	29	111	13	212	42
	%	77.69	22.31	89.52	10.48	83.46	16.54
	Proportion test	$p < 0.0001$		$p < 0.0001$		$p < 0.0001$	

Panel B. Subject choice in Treatment II (with added information)

H0: A = B		**Men**		**Women**		**Men + Women**	
Country		**A**	**B**	**A**	**B**	**A**	**B**
USA	Number	86	164	83	167	169	331
	%	34.40	65.60	33.20	66.80	33.80	66.20
	Proportion test	$p < 0.0001$		$p < 0.0001$		$p < 0.0001$	
India	Number	121	129	43	54	164	183
	%	48.40	51.60	44.33	55.67	47.26	52.74
	Proportion test	$p = 0.613$		$p = 0.267$		$p = 0.315$	
USA + India	Number	207	293	126	221	333	514
	%	41.40	58.60	36.31	63.69	39.32	60.68
	Proportion test	$p < 0.0001$		$p < 0.0001$		$p < 0.0001$	

Concerning Panel B, investors opting for Book B when compared to those choosing Book A in *Treatment II* were significant for all participants from the US but not significant for those of India. Although investment shifts experienced from *Treatment I* to *Treatment II* also favored Book B for the case of India, at the end, the number of investors supporting both books was not significantly different. Thus, we observed the prevalence of peers over experts at influencing investors' beliefs of Indian subjects but not as strongly as for the case of US subjects.

5. Conclusions

In order to analyze how electronic word of mouth (eWOM) influences the investment decision in reward-based crowdfunding platforms, it was conducted an economic experiment using Amazon Mechanical Turk (AMT), an internet-based crowdsourcing platform. According to results, eWOM influences investors' beliefs and, in so doing, modifies initial investment decisions. Additionally, recommendations of other buyers were found to be more influential than those of experts. It has practical implications for those seeking funding in crowdfunding platforms given that, in the light of these results, it becomes more important to encourage positive recommendations by previous buyers rather than excessively relying in experts' appraisal. Even though marketing campaigns often rely on famous experts as a way to increase sales and consumers' loyalty as well as encourage investors to contribute funds, current research puts the spotlight on peers for their power to influence others who share similar aims and motivations.

Further research should try to model the influence of eWOM over backers' investment decision in crowdfunding platforms building on theories such as signaling and herding behavior. It should also try to clarify if backers behave rationally based on observational learning or simply mimic others' behavior.

Ultimately, we tested AMT as a tool for experimental and behavioral research allowing for an instant access to a large and culturally diverse subject pool. Experimental research is usually constrained by costly laboratory-based experiments with small pools of subjects. AMT helps to overcome this issue by widening the subject pool that can be recruited both rapidly and inexpensively through Internet.

Acknowledgments

Irene Comeig acknowledges financial support from the Spanish Ministry of Economy through grant number ECO2016-75575-R.

References

Agrawal, A., Catalini, C., & Goldfarb, A. (2014). Some simple economics of crowdfunding. *Innovation Policy and the Economy*, 14, 63-97.

Belleflamme, P., Lambert, T., & Schwienbacher, A. (2014). Crowdfunding: Tapping the right crowd. *Journal of Business Venturing*, 29, 585-609.

Bi, S., Liu, Z., & Usman, K. (2017). The influence of online information on investing decisions of reward-based crowdfunding. *Journal of Business Research*, 71, 10-18.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1), 3-5.

Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods*, 46, 112-130.

Cosh, A., Cumming, D., & Hughes, A. (2009). Outside Entrepreneurial Capital. *The Economic Journal*, 119, 1494-1533.

Erkan, I., & Evans, C. (2016). The influence of eWOM in social media on consumers' purchase intentions: An extended approach to information adoption. *Computers in Human Behavior*, 61, 47-55.

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's Mechanical Turk. *Computers in Human Behavior*, 29, 1749-1754.

Huang, J.H., & Chen, Y.F. (2006). Herding in Online Product Choice. *Psychology & Marketing*, 23(5), 413-428.

Johnson, D. R., & Borden, L. A. (2012). Participants at Your Fingertips: Using Amazon's Mechanical Turk to Increase Student–Faculty Collaborative Research. *Teaching of Psychology*, 39(4), 245-251.

Lee, M., & Youn, S. (2009). Electronic word of mouth (eWOM). How eWOM platforms influence consumer product judgement. *International Journal of Advertising*, 28(3), 473-499.

Mason, W., & Siddharth, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavioral Research Methods*, 44(1), 1-23.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867-872.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), 411-419.

Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, 13(10), 420-428.

Schwienbacher, A. (2018). Entrepreneurial Risk-Taking in Crowdfunding Campaigns. *Small Business Economics*, 51(4), 843-859.

Simonsohn, U., & Ariely, D. (2008). When Rational Sellers Face Nonrational Buyers: Evidence of Herding on eBay. *Management Science*, 54(9), 1624-1637.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11512

An algorithm to fit conditional tail expectation regression models for vehicle excess speed in driving data

Albert Pitarque, Montserrat Guillen
Riskcenter UB, Spain.

Abstract

An algorithm to fit regression models aimed at predicted the average responses beyond a conditional quantile level is presented. This procedure is implemented in a case study of insured drivers covering almost 10,000. The aim is to predict the expected yearly distance driven above the posted speed limits as a function of driving patterns such as total distance, urban and night percent driven. Gender and age are also controlled. Results are analyzed for the median and the top decile. The conclusions provide evidence of factors influencing speed limit violations for risky drivers and they are interesting to price motor insurance and implement road safety policies. The efficiency of the algorithm to fit tail expectation regression is compared to quantile regression. Computational time doubles for tail expectation regression compared to quantile regression. Standard errors are estimated via bootstrap methods. Further considerations regarding in-sample predictive performance are discussed. In particular, further restrictions should be imposed in the model specification to avoid prediction outside the plausible range.

Keywords: *Telematics; quantile regression; insurance; tail value-at-risk; traffic safety.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

The analysis of data collected from vehicles in motion is an emerging area in transportation research. The reason for its growing interest is the possibility to obtain safety improvements on the road and to develop new ways to calculate motor insurance prices. The aim of this paper is to propose new models for risk analysis. We present an algorithm that allows adjusting regression models for the tail expectation that are a natural generalization of quantile regression models. Unlike the classical linear model, which finds the effects of covariates on the mean of a response variable, quantile regression identifies the effects on the quantile of the response. Tail expectation regressions model conditional average responses above a given conditional quantile. In our case study, we show that quantile regression identifies risky drivers by modelling quantiles of distance driven yearly above the posted speed limits. The quantile order is fixed at high levels, such as 95%. We denote as c_τ the quantile at the level τ (τ between 0 and 1) of a variable response Y. By definition, the probability that Y is greater or equal to c_τ is equal to τ. Quantiles are used in areas such as finance, insurance and risk analysis, where they are usually referred to as τ – Value at Risk (VaR_τ). Another risk measure is the Expected Shortfall (ES_τ) also known as Conditional Tail Expectation (CTE_τ) or Tail Value at Risk ($TVaR_\tau$). It is defined as:

$$TVaR_\tau(Y) = E(Y|Y > c_\tau). \tag{1}$$

Quantile regression and tail expectation regression specify $VaR_{\tau\tau}$ and $TVaR_\tau$, respectively, as a linear combination of regressors.

2. Methodology

The starting point for this work is quantile regression. Quantile regression is an extension of the linear regression that is especially interesting when the response variable has asymmetry, for instance when there is a substantial difference between the conditional mean and the conditional median. As it is widely known, the median is robust to the presence of outliers, while the mean is not. Koenker and Bassett (1978) proposed an optimization framework to fit quantile regressions. Here, a new procedure to estimate the tail expectation model is presented and it is implemented in open source software R.

A classical linear regression model is represented as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \dots \beta_k X_{ki} + \varepsilon_i, \tag{2}$$

where Y_i is the response variable for the i^{th} individual (i = 1, ... , n), X_{ji} represents the value of the i^{th} observation of explanatory variable j (j = 1, ... , k) and β_j is the j^{th} parameter. The i^{th} linear predictor is defined as $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \dots \beta_k X_{ki}$. The error term, ε_i, is the part of the response variable that cannot be explained by the covariates. Parameter β_0 is

known as the intercept and it is usually included in the model, so that it can be assumed that the error term has expectation equal to zero. Model (1) is usually estimated by ordinary least squares (OLS), i.e. by minimizing the sum of squared residuals:

$$\hat{\beta} = arg \min_{\beta} \sum_{i=1}^{n} f_i(\beta), \quad (3)$$

where $f_i(\beta) = (Y_i - X_i\beta)^2$ represents the difference between the observed response and the linear predictor.

Quantile regression assumes that the quantile at level τ of the response equals a linear combination of the regressors:

$$VaR_\tau(Y_i|X_{j1}, \dots, X_{ji}) = \beta_0^\tau + \beta_1^\tau X_{1i} + \beta_2^\tau X_{2i} + \cdots + \beta_k^\tau X_{ki}. \quad (4)$$

Coefficient estimates are obtained as follows (see Koenker and Bassett, 1979; Koenker and Machado 1999):

$$\widehat{\beta^\tau} = arg \min_{\beta} \sum_{i=1}^{n} [\rho_i^\tau (Y_i - X_{ij}\beta_j)]. \quad (5)$$

where ρ_i^τ represents a loss function of the quantile, which is equal to τ when $Y_i - X_i\beta$ is greater or equal than 0 and τ-1, otherwise. The standard deviation of the estimated coefficients can be calculated following the bootstrap method (Chernick, 2011; Hestenberg, 2011).

The specification of tail expectation regression is defined as:

$$TVaR_\tau(Y_i|X_{j1}, \dots, X_{ji}) = \beta_0^\tau + \beta_1^\tau X_{1i} + \beta_2^\tau X_{2i} + \cdots + \beta_k^\tau X_{ki}. \quad (6)$$

Acerbi and Szekely (2014) recently proposed a loss function to estimate the conditional tail expectation using the quantile. Despite developing this method theoretically, these authors did not consider a linear predictor. In the field of risk analysis, databases are large. This is the reason why we focus studying the optimization underlying the estimation procedure is of outmost interest. Computational time remains a challenge.

3. Data

Information about different characteristics of 9,614 car drivers was collected during 2010 by an insurance company, using a telematics device. Driving data measure patterns of vehicles in motion such as distance driven, vehicle speed, time of the day, and zone (urban versus nonurban). For privacy reasons, GPS localization data are not recorded. A definition of the variables is presented in Table 1. Drivers are aged between 18 and 35 years, because the insurance company offered a pay-as-you drive motor policy only to young drivers. Boucher et al. (2017) studied the transformation of the risk factors with the same dataset; Ayuso et al.(2016a, 2016b) compared the driving patterns between male and female

drivers; Guillen et al. (2019) proposed new methods to calculate the price of motor insurance. Pitarque et al. (2019) used quantile regression to analyse risk of having an accident.

Table 1. Definition of the variables in the insurance dataset (9,614 observations in 2010).

Variable	Description
Toler_km	Total number of kilometres driven exceeding the posted limit
lnKm	Logarithm of the total of kilometres driven
P_urban	Percentage of kilometres driven in urban areas
P_night	Percentage of kilometres driven during the night
Age	Age of the driver at 1st of January, 2010
Male	Gender of the driver (1 = male, 0 = female)

A descriptive analysis of the data is presented in Table 2. Skewness equal to 3.64 is one of the most relevant features of total distance driven above the posted speed limits during one year. This means that while most drivers have low levels of excess speeding, a few of them present large values. However, it is necessary to consider total driving distance, urban driving and night driving to extract conclusions.

Table 2. Descriptive statistics in the insurance dataset (9,614 observations in 2010).

	Mean	Median	Minimum	Maximum	Standard deviation	Skewness
Toler_km	1398.21	689.23	0.00	23500.19	1995.37	3.64
lnKm	9.27	9.37	-0.37	10.96	0.75	-1.87
P_urban	26.29	23.39	0.00	100.00	14.18	1.03
P_night	7.02	5.31	0.00	78.56	6.13	1.68
Age	24.78	24.63	18.11	35.00	2.82	0.11

4. Results

A simple quantile regression with only one explanatory variable is adjusted to model the percentage of kilometres driven above the speed limit with $\tau = 0.9$ as a function of the percentage of kilometres driven in urban areas. The tail expectation regression is also fitted. Parameter estimates are not displayed for brevity. The results are shown graphically in

Figure 1. Quantile regression at the 0.9 level indicates that when there is an increase of 1% in the percentage of kilometres driven in urban areas, the Value at Risk of the percentage of kilometres driven above the speed limit decreases by 0.35% and the average beyond the quantile level decreases 52 basis points.

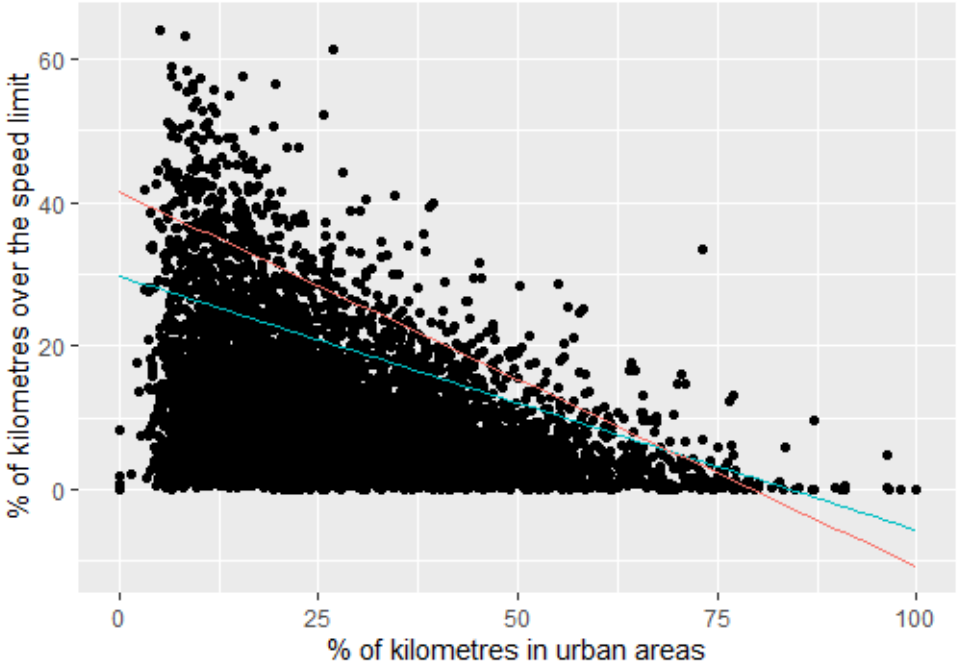


Figure 1. Graph of the relation between the percentage of kilometres driven above the speed limit and the percentage of kilometres driven in urban areas in the insurance dataset. Blue line represents a 90% quantile regression line and red line represents the a 90% tail expectation regression.

In the multivariate case, the total number of kilometres driven above the speed limit as the response variable is analysed for quantile levels $\tau = 0.5$ (median) and $\tau = 0.9$ (upper decile). A linear regression model is also estimated to compare the coefficient estimates.

Coefficient and standard deviation estimates are calculated using the *quantreg* package of R (Koencker et al., 2019). Standard errors were computed from 3.000 replications with samples of the same length as the original sample with replacement, so that a comparison between models can be analyzed. Table 3 presents results for the linear regression, the quantile regression and the tail expectation regression together with the goodness-of-fit statistic. As in the univariate case extrapolation of the linear specifications can produce abnormalities such as negative predictions or values of the conditional tail expectation lower than its corresponding quantile level. A summary is reported in Table 4.

Table 3. Models results of linear regression (OLS), quantile regression (VaR) and tail expectation regression (TVaR) for quantile levels $\tau = 0.5$ and $\tau = 0.9$ in the insurance dataset. In parenthesis, the standard errors of the estimated coefficients.

Variable	OLS	$VaR_{0.5}$	$TVaR_{0.5}$	$VaR_{0.9}$	$TVaR_{0.9}$
Intercept	-8082.51	-4496.53	-11708.92	-6418.11	-14068.39
	(309.95)	(186.02)	(843.57)	(742.98)	(3505.13)
lnKm	1064.51	597.60	1588.38	1074.66	2229.62
	(26.51)	(19.32)	(86.59)	(64.46)	(364.14)
P_urban	-21.87	-9.19	-39.72	-39.59	-86.08
	(1.39)	(0.62)	(2.16)	(2.34)	(7.14)
P_night	7.54	5.41	11.99	21.76	26.56
	(2.93)	(1.82)	(6.10)	(9.80)	(19.21)
Age	-1.13	-2.56	0.96	5.16	7.71
	(6.26)	(3.26)	(11.09)	(15.24)	(37.13)
Male	328.01	206.76	528.84	574.08	913.63
	(35.89)	(19.01)	(66.51)	(103.97)	(223.48)
R^2	0.25	0.14	0.17	0.20	0.49

Table 4. Percentage of cases where the predicted TVaR is lower than the predicted VaR and percentage of cases where the predicted TVaR is negative in the insurance database. Two quantile levels are considered $\tau = 0.5$ and $\tau = 0.9$.

% $TVaR_{0.5} < VaR_{0.5}$	8.20%
% $TVaR_{0.5} < 0$	7.41%
% $TVaR_{0.9} < VaR_{0.9}$	6.48%
% $TVaR_{0.9} < 0$	3.60%

The implementation of a routine to estimate the coefficients for the tail expectation regression can be compared with the VaR regression computation. An evaluation of computational time is presented in Table 5. The difference between TVaR regression and VaR regression is about double time both for the parameter estimates and the standard error. In both cases, the parameter estimates are obtained in less than 0.2 seconds for our working sample of almost 10 thousand cases and six coefficients. The most relevant result

is the time needed to compute the standard errors, which is quite low given the number of replicates. The quantile level did not affect computational time required.

Table 5. Computational time comparison in our case study.

Output generated	Computational time
Estimation of the VaR coefficients	0.088 seconds
Estimation of the standard deviation of the VaR coefficients	2.618 minutes
Estimation of the ES coefficients	0.175 seconds
Estimation of the standard deviation of the ES coefficients	5.410 minutes

5. Conclusions

An innovative method that generalizes quantile regression in order to study risky drivers was implemented. The study was done using a database containing approximately 10,000 observations, which contain a highly skewed response variable. This is a typical feature of risk analysis problem settings. In the case of the bivariate regression, the results show that the percentage of kilometres driven in urban areas influences the risk of exceed speed limits. In particular, each additional percent point driven in an urban area reduces the highest decile of the percentage of distance driven above the speed limits by 0,35% This decrease is emphasized in the case of the tail expectation where an increase of 1% in the percentage of kilometres driven in urban areas reduces 52 basis points the expected percentage of kilometres driven above the speed limit, for those drivers that are in the top decile.

In the multivariate case similar conclusions are drawn from quantile regression and tail expectation regression for quantile levels 0.5 and 0.9. Some problems arose when applying the models for an "in-sample" prediction exercise. In a few cases, the tail expectation was lower than the value provided by the quantile, or even negative. This could be a result of the simplicity of the linear specification and further research should be carried out to develop possible solutions to this issue. Despite those problems, the computational time of the estimation procedure to obtain the coefficient estimates is low, so the routine for the tail expectation regression that was created here is not excessively slow. The computational time for the standard errors is also relatively low, taking into account that the bootstrap method iterates the estimation in a large number of sample replicates.

For future studies, other methods to calculate the standard errors of the coefficient estimates should be investigated so that computational effort does not increase too much with sample size. Specially with the bootstrap method, there are currently several possible alternatives

that seem suitable to our problem. Another area for further analysis is larger datasets and tuning the parameters of the bootstrap method to estimate coefficients and standard errors in a reasonable computational time window.

References

Acerbi, C., & Szekely, B. (2014). Back-testing expected shortfall. *Risk*, 27(11), 76-81.

Ayuso, M., Guillen, M., & Pérez-Marín, A. M. (2016a). Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies*, 68, 160-167.

Ayuso, M., Guillen, M., & Pérez-Marín, A. M. (2016b). Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2), 1-10.

Boucher, J-P., Côté, S., & Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, 5(4), 54. https://doi.org/10.3390/risks5040054

Chernick, M. R. (2011). Bootstrap methods: A guide for practitioners and researchers (Vol. 619). John Wiley & Sons.

Guillen, M., Nielsen, J. P., Ayuso, M., & Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, 39(3), 662-672.

Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6), 497-526.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50.

Koenker, R., & Machado, A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448), 1296-1310.

Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P., & Ripley, B. D. (2019). Package 'quantreg'. https://cran.r-project.org/web/packages/quantreg/

Pitarque, A., Pérez-Marín, A. M., & Guillen, M. (2019). Quantile regression as a starting point in predictive risk models. *Anales del Instituto de Actuarios Españoles, 4ª época*, 25, 2019 /101-117

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11514

Regression scores to identify risky drivers from braking pulses

Shuai Sun[1,2], Jun Bi[1], Montserrat Guillen[2], Ana M. Pérez-Marín[2]
[1]Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, China, [2]Riskcenter, Universitat de Barcelona, Spain.

Abstract

Driving data record information on style and patterns of vehicles that are in motion. These data are analysed to obtain risk scores that can later be implemented in insurance pricing schemes. Scores may also be used in on-board sensors to create risk alerts that help drivers to keep up with safety margins. Regression methods are proposed and a prototype real sample of 253 drivers is analysed. Conclusions are drawn on the mean number of brake pulses per day as measured within 30 seconds time-intervals. Linear and logistic regressions serve to construct a label that classifies drivers. A novel factor based on the driving range that is defined from geo-localization improves the results considerably. Driving range is expressed as measures the diagonal of a rectangle that contains the furthest North-South versus East-West weekly vehicle trajectory. This factor shows that frequent braking activity is negatively related to the square of driving range.

Keywords: *Telematics; logistic regression; insurance; risk measures; traffic safety.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

The internet of vehicles is a new area providing lots of opportunities to develop big data applications before self-driving cars are fully available. In this paper, we analyze a set of 253 drivers that were monitored over one week. Data were collected every thirty seconds, including the position, speed, acceleration and the engine's revolutions per minute. No information on accident was available. We propose a new way to design driving risk alerts based on these data. Our predictive risk scores can serve as inputs to improve driving habits and to calculate insurance premiums by insurers in the Internet of Vehicle (IoV) environment. We also suggest that on-board vehicle telematics should encompass personalized risk-related alerts in their internal architecture.

Risk analysis in motor insurance or accident prevention usually studies traffic collisions (Handel et al, 2014). Some papers relate driving patterns or braking to accident risk (Jourbert et al., 2016; Guillen et al., 2019), as braking usually occurs before an accident happens, or before an accident is avoided.

Generalized linear models, have been used to predict the probability of a traffic accident or, alternatively, the expected frequency of insurance claims. As such, this is the main technique implemented by insurance companies to calculate the expected yearly number of claims and average cost, from which the basic premium prices are obtained (see, Jin et al., 2018; Verbelen et al., 2018; Ma et al., 2018). Paefgen et al. (2014) compared the performance of various machine learning methods, such as logistic regression, neural network, and decision tree classifiers, in driving risk prediction and insurance pricing. The interpretability of logistic regression models has made this method the outstanding technique to calculate risk scores. In many countries premium calculation is regulated and, as a result, the authorities prefer methods that are not black boxes (see, also Pesantez-Narvaez et al., 2019).

2. Data

Data used in our study were obtained from an IoV information service provider in China. Each vehicle in our database has a telematics box (T-box), including a GPS sensor, a vehicle condition sensor, and a wireless transmission unit. When the vehicle is turned on, data get recorded second-by-second and then they are aggregated on the device level to reduce costs of data transmission and storage. Every 30 seconds, the T-box transmits the latest piece of data to a central database. When the vehicle is turned off, the on-board device automatic restarts every 30 minutes and transfers a bunch of data to the base station.

The total number of effective vehicle data files is 253. Our statistical learning has to deal with unique vehicle identification, time-varying GPS trajectory data, and abundant vehicle

condition information. A summary of the per-vehicle averages can be found in Table 1. We also show in Figure 1 a Google Maps example of car travel trajectories in a Chinese region.

Table 1. The basic descriptive statistics of the variables in the driving data set 2019.

Variable	Definition	Mean	Standard Deviation	Minimum	Median	Maximum
Brakes	Brake times with speed>40km/h	1540.194	1266.109	26.000	1162.000	6633.000
Accelerator	Mean of acceleration pedal position (%)	19.640	7.313	0.185	20.124	39.480
Distance	Cumulative driving distance (Km)	2211.046	1578.700	17.140	1975.570	7163.830
Speed	Mean of speed (Km/h)	36.076	15.225	1.187	36.123	66.819
RPM	Mean of revolutions per minute	997.390	178.219	232.263	983.173	1622.257
Range	Range of driving (geographical units)	3.050	3.334	0.013	1.706	14.593

Source: Own calculations

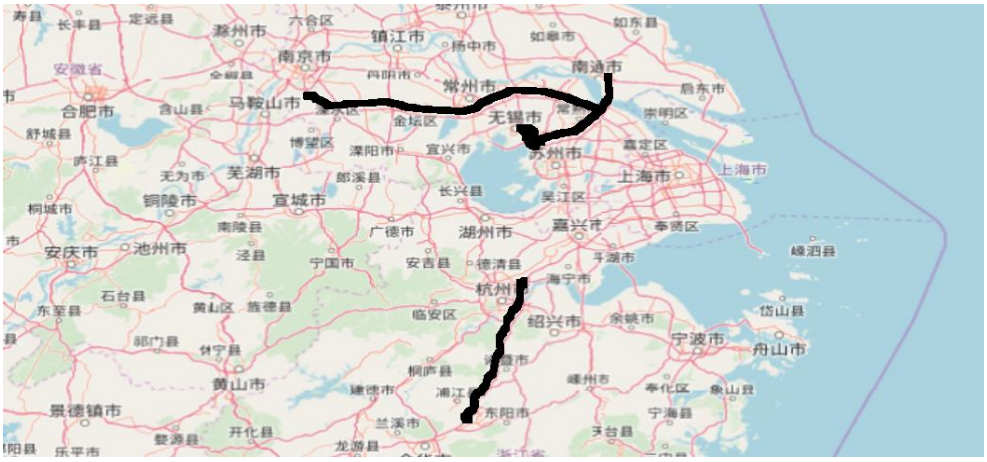


Figure 1. This map plots tow trajectories of monitored cars in July 2019.

Figure 2 presents the development of distance driven by one car in the upper plot. In the middle plot, there is a graph of the fuel consumption accumulated over time and finally, the speed and brake pulses are shown in the lower plot. For all cars in the sample, a daily average and variance was computed.

In addition, a new measure was defined to capture the driving pattern regarding the fact that a driver always stays in the same region. Driving range has a major influence on the analysis of driving risk. Whenever drivers brake or accelerate, there is either traffic congestion or moving requirements in limited area. However, this behavior must be relative to the driving radius. Indeed, a large number of brakes or accelerations for someone who drives longer distances, in a large radius from the starting home point may indicate that there may be safety hazards, compared to someone staying in a short circle distance. Driving range was calculated based on the available GPS trajectory as follows:

$$Range = \sqrt{(lon_{max} - lon_{min})^2 + (lat_{max} - lat_{min})^2}$$

where lon_{max} and lon_{min} represents the maximum and minimum observed Longitude value, lat_{max} and lat_{min} represents the maximum and minimum observed Latitude value.

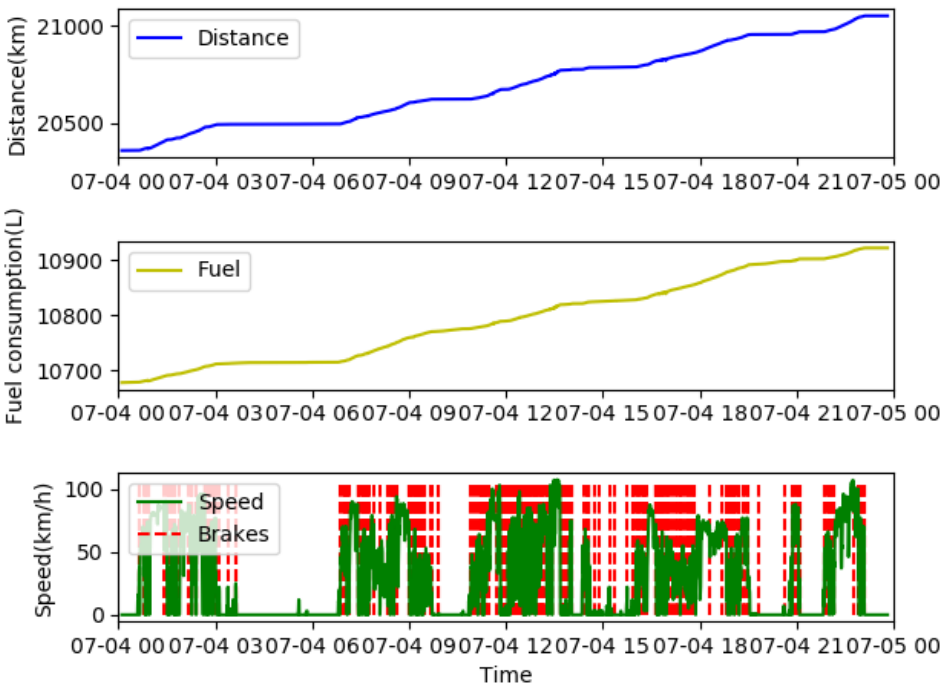


Figure 2. Example of accumulated distance driven (upper), fuel consumption (middle) and speed/breaking activity measured for a vehicle in July 2019.

3. Methods and results

OLS regression and logistic regression were respectively estimated by taking "Brakes" as the dependent variable. The variables derived from the multiplication of two pairs are taken as independent variables to reflect the interaction between variables.

In the logistic regression model variable "Brakes" was dichotomized as follows. Brake levels above the median were identified as events, while those below the median were identified as non-events. The stepwise regression results after bidirectional elimination are shown in Table 2. The goodness-of-fit (pseudo-)R^2 statistic is 34% and 27% for the linear and logistic regression, respectively.

Table 2. Coefficient estimates and P-values for linear regression (left) and logistic regression (right) in the driving data set 2019.

	OLS		Logistic	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept	-359.5976	0.173	-2.1345	0.000
Distance	1.4935	0.003	0.0016	0.000
Speed	-85.1223	0.045		
Accelerator	99.5726	0.014		
Distance^2	8.574e-05	0.012		
Speed^2			-0.0029	0.001
RPM^2			-5.309e-06	0.005
Accelerator^2			-0.0094	0.024
Distance*Speed	-0.0180	0.000		
Distance*RPM	-0.0018	0.006		
Distance*Accelerator	0.0555	0.000		
Range^2	-8.7141	0.007		
Range*RPM	0.2533	0.002		
Range*Accelerator	-7.2562	0.018		
Speed*RPM	0.1516	0.003	0.0001	0.045
Speed*Accelerator	-3.1860	0.002		
RPM*Accelerator	-0.0818	0.019	0.0004	0.017

Source: Own calculations.

A comparison between intuitive judgement and in-sample prediction shows that there is a mismatch between the conditional scores produced by the multivariate models, i.e. when taking into account the driver's information, and the intuitive judgment that is solely based on the univariate analysis of brake pulses.

This is easily seen in Figure 3, where there is a comparison between the predicted scores provided by the models (vertical axis) and the value of the observed brakes (horizontal axis).

A risky driver has a predicted score that is lower than his observed braking value. A non-risky driver has a predicted score that is higher than his real observed breaking value. Alternatively, the median of the scores and the median of the observed braking values are used as classifiers as shown in Figure 3. Drivers in the right bottom box are the ones that would be identified by our models as risky drivers. The two models produce slightly different results in terms of identified risky drivers.

Sensors should dynamically react to large values of brake pulses. The regression line or the logistic curve should act as the fundamental pieces of personalized alert systems. For example, a driver that has brake some of pulses equal to 300, but whose risk score predicts a total of 200, should be warned because his level of braking activity is above the risk limit predicted by the model.

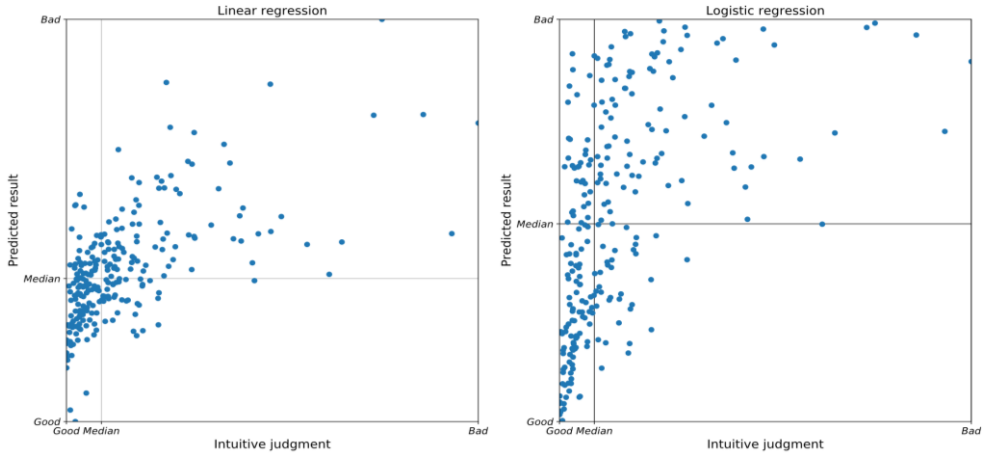


Figure 3. Comparison of predicted versus observed scores for "Brakes" in the telematics data set. Left plot corresponds to linear regression and right plot corresponds to logistic regression.

The points in Figure 3 indicate that the response variable has some outlying observations and that it is right skewed. The difference between the intuitive judgement, which corresponds to the purely observed values (horizontal axis) and the predicted scores (vertical axis) is that the later take into consideration the driving characteristics included in

the model. Those drivers exhibiting especially good habits and observed values below the predicted scores should be rewarded by the insurance company due to good driving habits. Insurance companies may deny access to insurance for drivers who cause extreme driving hazards in order to diminish claims.

Scores can also serve to reveal problems in the vehicle sensors. We argue that some on-board devices suffer quality deterioration over time. Predictive scores that are systematically above the levels of observed braking activity may indicate that the sensor fails to transmit the true driving activity correctly. Driving data providers should be aware that scoring drivers benefits traffic safety, insurance companies and their own business quality control.

4. Conclusions

Accident risk analysis is difficult because collisions and crashes seldom occur. The inspection of IoV data even if there is no information on motor accidents can be done by comparing drivers' ratings and observed patterns. Basic machine learning models were used to classify observations and to identify risky clusters of drivers in the sample. The mean of the braking pulses when the vehicle exceed 40 Km/h was used as a response variable and it was also dichotomized to reveal an association with other driving factors. This solution is promising for insurers and even car manufacturers that design new safety procedures and gadgets. Data analysis of a big source of information when accident data for vehicles were not available is still valuable to produce relevant scores. The relationship between accident risk has been found in previous studies. So, the level of braking pulse intensity can be related positively with proportionally higher insurance prices (Bian et al., 2018; Carfora et al., 2019; Tselentis, 2016 and 2017).

The linear relationship between the response variable and the covariates is a limitation of the linear regression model. Further analysis of more flexible specifications is recommended. Pérez-Marín and Guillen (2019) showed that excess speed is one of the main factor influencing driving risk, however the results obtained for this dataset indicate that mean speed is inversely related to braking, but positively related to braking when interacting with engine RPM. The higher the speed, the more acceleration action is required. Moreover, the higher the RPM, the more braking action is need to reduce speed, and so the greater the driving risk.

Driving range was not informative by itself, but it did have a substantial influence when combines with other factors.

Some of the limitations of this case study are related to the model approaches. Other machine learning algorithms and classifying techniques should also be examined. Some

additional efforts remain in the research agenda. For example, geolocation information could be used to define driving zones (urban versus nonurban) and time stamps could be transformed into day and night driving percent. The volume of information contained in each vehicle daily file opens an opportunity to explore patterns that may improve driving habits and produce recommendations for safety on the road.

Funds

The paper was granted by the Fundamental Research Funds for the Central Universities 2019YJS091.

Acknowledgements

MG thanks ICREA Academia. The authors thank Fundación BBVA and the China scholarship council.

References

Bian, Y., Yang, C., Zhao, J. L., Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. *Transportation Research Part A: Policy and Practice*, 107, 20-34.

Carfora, M. F., Martinelli, F., Mercaldo, F., Nardone, V., Orlando, A., Santone, A., & Vaglini, G. (2019). A "pay-how-you-drive" car insurance approach through cluster analysis. *Soft Computing*, *23*(9), 2863-2875.

Guillen, M., Nielsen, J. P., Ayuso, M., & Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis,* 39(3), 662-672.

Handel, P., Skog, I., Wahlstrom, J., Bonawiede, F., Welch, R., Ohlsson, J., et al. (2014). Insurance telematics: Opportunities and challenges with the smartphone solution. *IEEE Intelligent Transportation Systems Magazine*, 6(4), 57-70.

Jin, W., Deng, Y., Jiang, H., Xie, Q., Shen, W., & Han, W. (2018). Latent class analysis of accident risks in usage-based insurance: Evidence from Beijing. *Accident Analysis & Prevention*, 115, 79-88.

Joubert, J. W., de Beer, D., de Koker, N. (2016). Combining accelerometer data and contextual variables to evaluate the risk of driver behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour*, 41, 80-96.

Ma, Y., Zhu, X., Hu, X., Chiu, Y. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice*, 113, 243-258.

Paefgen, J., Staake, T., Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, 61, 27-40.

Pérez-Marín, A. M., Guillen, M. (2019). Semi-autonomous vehicles: Usage-based data evidences of what could be expected from eliminating speed limit violations. *Accident Analysis and Prevention*, 123, 99-106.

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. *Risks*, *7*(2), 70.

Tselentis, D. I., Yannis, G., Vlahogianni, E. I. (2016). Innovative insurance schemes: pay as/how you drive. *Transportation Research Procedia*, 14, 362-371.

Tselentis, D. I., Yannis, G., Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. *Accident Analysis and Prevention*, 98, 139-148.

Verbelen, R., Antonio, K., Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5), 1275-1304.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11557

Pruned Wasserstein Index Generation Model and wigpy Package

Fangzhou Xie
Department of Economics, New York University, USA, Department of Economics, Rutgers University, USA.

Abstract

Recent proposal of Wasserstein Index Generation model (WIG) has shown a new direction for automatically generating indices. However, it is challenging in practice to fit large datasets for two reasons. First, the Sinkhorn distance is notoriously expensive to compute and suffers from dimensionality severely. Second, it requires to compute a full $N \times N$ matrix to be fit into memory, where N is the dimension of vocabulary. When the dimensionality is too large, it is even impossible to compute at all. I hereby propose a Lasso-based shrinkage method to reduce dimensionality for the vocabulary as a pre-processing step prior to fitting the WIG model. After we get the word embedding from Word2Vec model, we could cluster these high-dimensional vectors by k-means clustering and pick most frequent tokens within each cluster to form the "base vocabulary". Non-base tokens are then regressed on the vectors of base token to get a transformation weight and we could thus represent the whole vocabulary by only the "base tokens". This variant, called pruned WIG (pWIG), will enable us to shrink vocabulary dimension at will but could still achieve high accuracy. I also provide a wigpy[1] *module in Python to carry out computation in both flavors. Application to Economic Policy Uncertainty (EPU) index is showcased as comparison with existing methods of generating time-series indices.*

Keywords: *Wasserstein Index Generation Model (WIG); Lasso Regression; Pruned Wassersteinn Index Generation (pWIG); Economic Policy Uncertainty Index (EPU).*

[1] https://github.com/mark-fangzhou-xie/wigpy

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

Recently, the Wasserstein Index Generation model (Xie, 2020) was proposed to generate time-series sentiment indices automatically. There have been several methods (Azqueta-Gavaldón, 2017; Baker, Bloom, & Davis, 2016; Castelnuovo & Tran, 2017; Ghirelli, Pérez, & Urtasun, 2019) proposed to generate time series sentiment indices, but, to the best of my knowledge, WIG is the first automatic method to produce sentiment indices completely free of manual work.

The WIG model runs as follows. Given a set of documents, each of which is associated with a timestamp, it will first cluster them into several topics, shrink each topic to a sentiment score, then multiply weights for each document to get document sentiment, and then aggregate over each time period. However, its computation on large dataset come with two challenges: (1) the calculation for Sinkhorn algorithm suffers from its notoriously computational complexity and the computation will soon become prohibitive; (2) this Optimal Transport-based method requires to compute a full $N \times N$ matrix, where N is the size of vocabulary, and it will become impossible to fit this distance matrix into memory after some threshold. Therefore, I propose a pruned Wasserstein Index Generation model (pWIG) to reduce dimensionality of vocabulary prior to fitting into the WIG model. This variant could represent the whole corpus in a much smaller vocabulary and then be fit in any memory-limited machine for the generation of time-series index. What is more, I also provide the *wigpy*[2] package for Python that could perform both version of WIG computation.

This paper first contributes to the EPU literature and tries to provide better estimations of that seminal time-series indices automatically. This article also relates itself to the new area of Narrative Economics (Shiller, 2017), where we could extract time-series sentiment indices from textual data, and thus provide a better understanding of how do narratives and sentiments relate to our economy.

2. Pruned Wasserstein Index Generation Model

We first review the original WIG model.

2.1. Review of Wassserstein Index Generation model

A major component of WIG model is the Wasserstein Dictionary Learning (Schmitz et al., 2018). Given a set of document $Y = [y_m] \in \mathbb{R}^{N \times M}$, each doc $y_m \in \Sigma^N$ is associated with a timestamp and N, M are length of dictionary and number of documents in corpus, respectively. Our first step is to cluster documents into topics $T = [t_k] \in \mathbb{R}^{N \times K}$, where

[2] https://github.com/mark-fangzhou-xie/wigpy

$K \ll M$, and associated weights $\Lambda = [\lambda_m] \in \mathbb{R}^{K \times M}$. Thus, for a single document y_m, we could represent it as $y_m \approx t_k \lambda_m$. Documents and topics lie in N-dimensional simplex and are word distributions. Another important quantity for computing WIG, is the cost matrix $C^{N \times N}$ and $C_{ij} = d^2(x_i, x_j)$, where each $x_i \in \mathbb{R}^{1 \times D}$ is the D-dimensional word embedding vector for the i-th word in the vocabulary. In other words, matrix C measures the "cost" of moving masses of words, and now we can proceed and define the Sinkhorn Distance.

Definition 1 (Sinkhorn Distance).

Given discrete distributions $\mu, \nu \in \mathbb{R}_+^N$, and C as cost matrix,

$$S_\varepsilon(\mu, \nu; C) \coloneqq \min_{\pi \in \Pi(\mu,\nu)} \langle \pi, C \rangle + \varepsilon \mathcal{H}(\pi)$$

$$s.t. \;\; \Pi(\mu, \nu) \coloneqq \{\pi \in \mathbb{R}_+^{N \times N}, \pi 1_N = \mu, \pi^T 1_N = \nu\},$$

where $\mathcal{H}(\pi) \coloneqq \sum_{ij} \pi_{ij} \log(\pi_{ij} - 1)$, negative entropy, and ε is the Sinkhorn regularization weight.

We could then set up the loss function and minimization problem as follows:

$$\min \sum_{m=1}^{M} \mathcal{L}\left(y_m, y_{S_\varepsilon}(T(R), \lambda_m(A); C, \varepsilon)\right),$$

$$s.t. \;\; t_{nk}(R) \coloneqq \frac{e^{r_{nk}}}{\sum_{n'} e^{r_{n'k}}}, \lambda_{nk}(A) \coloneqq \frac{e^{a_{km}}}{\sum_{k'} e^{a_{k'm}}}.$$

By this formula, we wish to minimize the divergence between original document y_m and the predicted (reconstructed) $y_{S_\varepsilon}(\cdot)$ given by Sinkhorn distance. Moreover, the constaints of this minimization problem considers *Softmax* operation on each of the columns of the matrices R and A, so that T and Λ will be (column-wise) discrete densities, as is required by the Sinkhorn distance.

For computation, we first initialize matrices R and A by drawing from Standard Normal distribution and then perform Softmax to obtain T and Λ. During training process, we keep track of computational graph and obtain the gradient $\nabla_T \mathcal{L}(\cdot\,; \varepsilon)$ and $\nabla_\Lambda \mathcal{L}(\cdot\,; \varepsilon)$ with respect to T and Λ. R and A are then optimized by Adam optimizer (Kingma & Ba, 2015) after each batch, and the automatic differentiation is done by PyTorch framework (Paszke et al., 2017).

After conducting Wasserstein Dictionary Learning on documents for clustering, the next step of WIG would be to generate time-series indices based on the topics. The model first reduces each topic vector t_k to a scalar by Singular Value Decomposition and then multiply the weight matrix to get document-wise sentiment score for the whole corpus. We then add up the scores for each month and then produce the final monthly index.

2.2. Pruned WIG (pWIG) Model

Although enjoying many nice theoretical properties (Villani, 2003), the computation for Optimal Transport has been known for its complexity. This burden has been eased by Cuturi (2013) and it has attracted much attention in machine learning community since then.

However, there are still two aspects that hindering our application to textual analysis. First of all, vocabulary will easily go to a very large one, and the computation for Sinkhorn loss will soon become prohibitive. Moreover, after passing a certain point, it not even possible to fit the distance matrix C into the memory, especially when considering the limited VRAM for GPU acceleration.[3]

I therefore propose the following procedure to reduce the vocabulary dimension and could avoid feeding the full vocabulary matrix into WIG model. It first clusters all word vectors by k-means clustering, and then selects a subset of tokens from each of the cluster to form "base tokens."[4] We could then use Lasso[5] to regress word vectors of all other tokens on the vectors of these "base tokens" to ensure sparse weight vector, which will have zero component on non-import features.

Formally speaking, we set up the following minimization problem for the k-means clustering:

$$argmin_{\mathcal{K}_1,\mathcal{K}_2,\cdots,\mathcal{K}_n} \sum_{k=1}^{K} \sum_{x\in\mathcal{K}_k} ||x-\mu_k||,$$

where μ_k is the mean of points in cluster $\mathcal{K}_k$ and $k \in \{1,\cdots,K\}$. We can certainly choose some most frequent tokens from each cluster to form a final subset whose length matches our desire.[6] By doing so, we also represent the whole vocabulary by the most representative tokens. The indices for these "base tokens" are collected in the index set,

---

[3] My configuration is Nvidia 1070Ti (8G). Under single precision, each digit will occupy 4 bytes, and, in my case, I can only fit, theoretically at most, a square matrix of dimension 44,721. I have a relatively small dataset from The New York Times and my vocabulary is of length 9437, but many NLP applications will have much more tokens than I do. In such a case, the WIG model will become infeasible.

[4] The number of tokens to be considered as "base tokens" is arbitrary, meaning that the compression ratio could potentially be made arbitrarily small. In other words, the researcher could choose such a number that the model can be fitted into the memory of her machine, regardless of the number of tokens she had for the corpus. And that is exactly the way why we need to compress the dictionary by "pruning" some non-important tokens.

[5] A similar approach (Mallapragada, Jin, & Jain, 2010) was proposed using group-Lasso to prune visual vocabulary, but in the area of image processing.

[6] A very simple choice would be $Word\ per\ Cluster = \frac{Maximum\ Vocabulary\ Length}{Number\ of\ Clusters}$.

$$\mathfrak{B} = \{b \in \{1, \cdots, N\} \mid x_b = 1\}.$$

Obviously, $\mathfrak{B}^C$ is also defined by excluding "base tokens" from the whole vocabulary. N is the size of vocabulary and x_b is the b-th token in the vocabulary.

Denote word vector for "base tokens" as v_b and others as v_o, we have

$$v_o = \sum_{b=1}^{B} \alpha_{o,b} v_b + \lambda \sum_{b=1}^{B} |\alpha_{o,b}|.$$

For each o, we will have a vector $\alpha_{o,b}$ of length B, where B is the dimension of "base vocabulary."

Previously in the WIG model, we obtain the word distribution for each single document y_m by calculating its word frequency, and that will give us a N-dimensional distribution vector. Here, in the pWIG variant, we replace the non-base tokens by weighted base-tokens and could thus represent the word simplex of documents in only B-dimensional spaces.

Now that we have successfully represent our dataset in s smaller vocabulary, we could proceed to define our distance matrix $C_{ij} = d^2(x_i, x_j)$, where $i, j \in \mathfrak{B}$. Here we have everything we need for the regular WIG model and we fit it using the shrinkage-transformed word distributions and distance matrix.

3. Numerical Experiments

3.1. wigpy *Package for Python*

To carry out the computation of WIG and pWIG model, I also provide the *wigpy* package under MIT license. Notice that the original WIG model is a new implementation, though part of the codes is modified from the codes of original WIG paper.

The main model is wrapped in the "WIG" class, where it contains a set of hyperparameters[7] to tune the model, and some parameters to control the behavior of preprocessing and Word2Vec training process.

Notice that the previous implementation of WIG model only supports hand-written Adam optimizer, and the optimization for document weights were optimized column-wise. In other words, each document will only be used to update the column of weight in matrix Λ for that given document. The new implementation wraps the whole model in PyTorch,

[7] For example, embedding depth ($emsize$), batch size ($batch\_size$), number of topics (num_topics), Sinkhorn regularization weight (reg), optimizer learning rate (lr), $L2$ penalty for optimizer ($wdecay$), $L1$/LASSO weight ($l1\_reg$), maximum number of tokens allowed by pWIG algorithm ($prune\_topk$).

providing many optimizers to choose by PyTorch optimizer class. What is more, each document will accumulate gradient and the whole Λ matrix will be updated all together.

3.2. Application to Generating Economic Policy Uncertainty Index (EPU)

To test for the pWIG model's performance, I run the model on the same dataset from the WIG paper. It consists of news headlines collected from The New York Times from 1980 to 2018. As I am implementing a new version of WIG, as provided by the *wigpy* module, I run the original WIG model and report its result as well.

I run both variants of WIG model separately, by calling wigpy package, to set for hyper-parameters by splitting training, evaluation, and testing data as 60%, 10%, and 30%, respectively.

For the original WIG, hyper-parameters are chosen as follows: depth of embedding $D = 50$, batch size $s = 32$, number of topics $K = 4$, learning rate for Adam $\rho = 0.001$, Sinkhorn regularization weight $\varepsilon = 0.1$; for the pWIG, depth of embedding $D = 50$, batch size $s = 64$, number of topics $K = 4$, learning rate for Adam $\rho = 0.001$, Sinkhorn regularization weight $\varepsilon = 0.08$.

I also report Pearson's and Spearman's correlation test on four set of automatically generated EPU indices (one LDA-based EPU (Azqueta-Gavaldón, 2017), one WIG-based EPU (Xie, 2020), and two flavor of WIG given by *wigpy* package in this paper), against the original EPU[8] (Baker et al., 2016).

Table 1. Pearson's and Spearman's correlation statistics[9]

EPU Flavor	Pearson's	Spearman's
LDA	77.48%	75.42%
WIG	80.24%	77.49%
WIG-wigpy	**80.53%**	**77.71%**
pWIG-wigpy	80.50%	77.64%

Apparently, as is shown in Table 1, all three WIG methods outperform LDA-based method by 3% in Pearson's test and more than 2% in Spearman's test. This fact has been established by the previous WIG paper. Moreover, if we compare results within three WIG-

[8] https://www.policyuncertainty.com/

[9] Since the LDA-based EPU was only available from 1989-2016, the test is performed using time-series indices within the same range.

related methods, this new implementation of original WIG in *wigpy* package shows better result than the previous implementation. The pruning method does not differ much from the new implemented WIG algorithm and is even better than the previous implementation of original WIG!

Table 2. Correlation statistics with other indices[10]

	VIX Pearson's	VIX Spearman's	Michigan Pearson's	Michigan Spearman's
WIG-wigpy	34.20%	19.56%	-56.40%	-49.38%
pWIG-wigpy	34.27%	19.82%	-56.45%	-49.62%

In Table 2, the correlation statistics between EPU generated by WIGs and two other indices: VIX and Michigan Consumer Confidence Sentiment index. As reported (Baker et al., 2016), EPU has a correlation of 0.58 between VIX and -0.742 between Michigan index. Since our objective is to produce a similar index of EPU, but using an automatic approach, we should expect our WIG-based EPU to have a similar relationship with these other two indices. This is indeed the case here, and we can certainly observe the positive and negative relationship when comparing the VIX and Michigan indices[11].

[10] Here I am comparing both flavors of WIG indices with VIX index and Michigan Consumer Sentiment index, using both Pearson's and Spearman's test. As VIX is only available up to 1986, and the WIG indices was generated up to 2018, I therefore take all the indices from 1986 to 2018 to perform the test. As usual, all indices are scaled to have mean 100 and unit standard deviation. Moreover, the correlation between two WIG indices is 99.86%.

[11] It may be confusing why the "sentiment index" generated by WIG models has a negative relationship with "Michigan Consumer Sentiment index," since both names contain "sentiment." However, there is a clear distinction of the usage of the same word in two different contexts. The famous Michigan index is expressed as the consumer confidence levels, and the higher the index, the more confident the consumers are. The word "sentiment", as used by WIG, is to capture the subjective information expressed in the texts. In the application of EPU, it is used to capture the intensity of opinions towards the uncertainty of policy, as conveyed by newspaper articles. It is very obvious that what it captures is negative feelings, and the higher the index, the more uncertain that people feel. In other words, although bearing the same word "sentiment" in their names, the underlying element is strikingly different and thus show a negative relationship between each other. Moreover, the WIG model does not limit its use in EPU. As soon as we apply the WIG models to other (textual) datasets, the meaning of "sentiment" will be changed accordingly. In total, the word "sentiment" used in WIG models is more versatile and should be distinguished from the usage as in the Michigan index.

3. Conclusion

This paper further extends the Wasserstein Index Generation (WIG) model, by selecting a subset of tokens to represent the whole vocabulary to shrink the dimension. The showcase of generating EPU has shown that the performance is retained while dimension being reduced. Moreover, a package, *wigpy*, is provided to carry out the computation of two variants of WIG.

References

Azqueta-Gavaldón, A. (2017). Developing news-based Economic Policy Uncertainty index with unsupervised machine learning. *Economics Letters*, *158*, 47–50.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, *131*, 1593–1636.

Castelnuovo, E., & Tran, T. D. (2017). Google It Up! A Google Trends-based Uncertainty index for the United States and Australia. *Economics Letters*, *161*, 149–153.

Cuturi, M. (2013). Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 2292–2300). Curran Associates, Inc.

Ghirelli, C., Pérez, J. J., & Urtasun, A. (2019). A new economic policy uncertainty index for Spain. *Economics Letters*, *182*, 64–67.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Mallapragada, P. K., Jin, R., & Jain, A. K. (2010). Online visual vocabulary pruning using pairwise constraints. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3073–3080.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., … Lerer, A. (2017). Automatic differentiation in PyTorch. *NIPS-W*.

Schmitz, M. A., Heitz, M., Bonneel, N., Ngolè, F., Coeurjolly, D., Cuturi, M., … Starck, J.-L. (2018). Wasserstein Dictionary Learning: Optimal Transport-Based Unsupervised Nonlinear Dictionary Learning. *SIAM Journal on Imaging Sciences*, *11*, 643–678.

Shiller, R. J. (2017). Narrative Economics. *American Economic Review*, *107*, 967–1004.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

Xie, F. (2020). Wasserstein Index Generation Model: Automatic generation of time-series index with application to Economic Policy Uncertainty. *Economics Letters*, *186*, 108874.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11560

Model degradation in web derived text-based models

Piet J.H. Daas[1,2], Jelmer Jansen[3]
[1]Division of Corporate services, IT and Methodology, Statistics Netherlands, the Netherlands, [2]Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands, [3]Faculty of Social Science, Radboud University Nijmegen, the Netherlands.

Abstract

Getting an overview of the innovative companies in a country is a challenging task. Traditionally, this is done by sending a questionnaire to a sample of large companies. For this an alternative approach has been developed: determining if a company is innovative by studying the text on the main page of its website. The text-based model created is able to reproduce the results from the survey and is also able to detect small innovative companies, such as startups. However, model stability was found to be a serious problem. It suffered from model degradation which resulted in a gradual decrease in the detection of innovative companies. The accuracy of the model dropped from 93% to 63% over a period of one year. In this paper this phenomenon is described and the data underlying it is studied in great detail. It was found that the combination of the inactivity of a subset of websites and changes in the composition of the words on company websites over time produced this effect. A solution for dealing with this phenomenon is presented and future research is discussed.

Keywords: *Innovation; Text analysis; Webscraping; Big data.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

In our modern world, more and more data are generated on the web and produced by sensors in the ever growing number of electronic devices surrounding us. These data have very interesting potential applications such as provide novel insights on the activities of companies (Gökk *et al.,* 2015), to inform policymakers (Höchtl *et al.,* 2015) and also for official statistics (Florescu *et al.,* 2014), especially when performed at large scale. However, extracting relevant and reliable information from Big Data sources in a reproducible way is not an easy task (Kitchin, 2015, Daas *et al.,* 2015).

In this paper we describe the results of our research on the development and application of text-based models for official statistics. The focus is on the development of a model that is able to detect innovative companies based on the text displayed on their website (Daas & van der Doef, 2020). Here, the occurrence of particular words and their co-occurrences with other words are used to determine if a company is innovative or not. The findings of the Community Innovation Survey (CIS) in the Netherlands are used for model development. The initial logistic regression model created had an accuracy of 93% and an F1-score of 93% on the test set (Daas & van der Doef, 2020). Kinne and Lenz (2019) demonstrated that the text on the websites of German companies could also be used to detect innovative companies. Their Deep Learning based model had an F1-score of 80%.

1.1. Goal of this study

During the work on Dutch company websites it was discovered that the performance of the initial model gradually decreased over time (more on that below). The model could no longer properly detect positive examples of innovation when exposed to more recently crawled data. The work described in this paper focusses on this issue, which very much resembles a phenomenon known as 'concept drift' in the Machine Learning world (Lu *et al.,* 2018). In this paper, we particularly focus on the underlying causes of the fact that the model produced less and less accurate results, to better understand what is exactly happening. Ultimate goal of the work is to assure the production of high quality output derived from new data sources collected at different points in time.

2. Concept drift

Certainly for the analysis of texts, models need to be used to measure the phenomenon the researcher is interested in. This indicates that the concept cannot be directly measured and thus one has to rely on the statistical properties of features in the text (Jo, 2019). When these properties change over time, the model-derived results are not stable (Zhang *et al.,* 2017). This phenomenon is known as concept drift (Lu *et al.,* 2018) which has attracted much attention in the Machine Learning community, especially by those that study so-

called data streams. Concept drift describes unforeseeable changes in the underlying distribution of streaming data over time. In general, four types of drift are discerned which are: i) Sudden drift, ii) Gradual drift, iii) Incremental drift and iv) Reoccurring concepts (Lu *et al.,* 2018, Fig 4). In the context of text-based models, reduction of the long-term stability has, for instance, been observed in news topic classification (Kim & Hovy, 2006) and event detection and tracing (Atefeh & Khreich, 2015).

2.1. Detecting innovation

The initial innovation detection model was developed on a set of 2,529 innovative and 2,236 non-innovative websites that all contained 10 or more words after processing (Daas & van der Doef, 2020). The processing steps were implemented in Python and were as follows. First, all script and style sections were removed from the scraped and parsed web pages, followed by language detection of the visible text. Since the majority of the pages were either written in Dutch or English, only those languages were discerned; i.e. any non-Dutch page was considered written in English. Subsequently, all words were converted to lower case and all punctuation marks, numbers, and all words with less than 3 characters were removed. Next, depending on the language detected, any words included in the stop words list for that language were removed. This was followed by stemming with the SnowballStemmer library. The words and language were subsequently used as features in model development using the well-known representation of frequency-annotated bag-of-words (Aggarwal, 2016). As already indicated above, ten words per website was the minimum amount of words considered for classification. For more details the reader is referred to Daas and van der Doef (2020).

The initial logistic regression model contained 180 words had an accuracy of 93% on the test set. But on freshly scraped data -from the same list of websites- 6 months after the original model was developed, the accuracy of the model was found to be 76% and after one year it was as low as 63% (Figure 1). As is clear from Figure 1, the accuracy slowly decreased over time, suggesting a gradual degradation of the concept measured. The big question here is ‘What exactly caused this decline?’. For this we will compare the first (t = 0) and last (t = 12) data sets in more detail.

2.2. Data sets compared

The study started with an initial set of 3,338 innovative and 2,876 non-innovative companies for which a website was found. As is clear from the numbers of websites finally used to create the model (see above), not all websites could be scraped and not all resulted in 10 or more words after processing. When comparing the companies for which a website was scraped and used in model building, it was found that these numbers were not identical

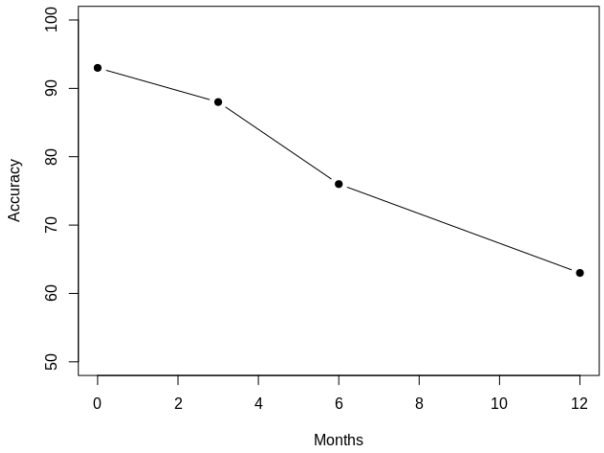


Figure 1. Accuracy of the innovation detection model over time on websites of companies included in the CIS survey scraped at different points in time. The model was developed at month 0. The average accuracy of the model, after 10-fold cross-validation, on the scraped webpages is shown.

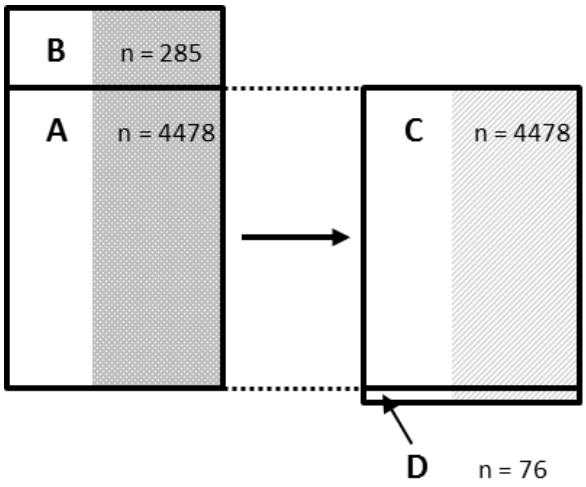


Figure 2. Graphical representation of the difference in the composition of the first (t = 0) and last (t = 12) data sets scraped. The data sets have 4,478 company web pages in common. For these the text extracted may differ which is illustrated by the different shades of gray.

in both data sets. This is illustrated in Figure 2. From this figure it is clear that the websites of 4,478 companies are included in both data sets. A total of 285 companies only occur in the first data set and, to our surprise, 76 only occur in the last data set. Comparison of the texts obtained after processing of the 4,478 websites included in both data sets revealed that these also differed. The similarity of the texts, expressed as the number of words in common in both versions of each website divided by the total number of unique words in

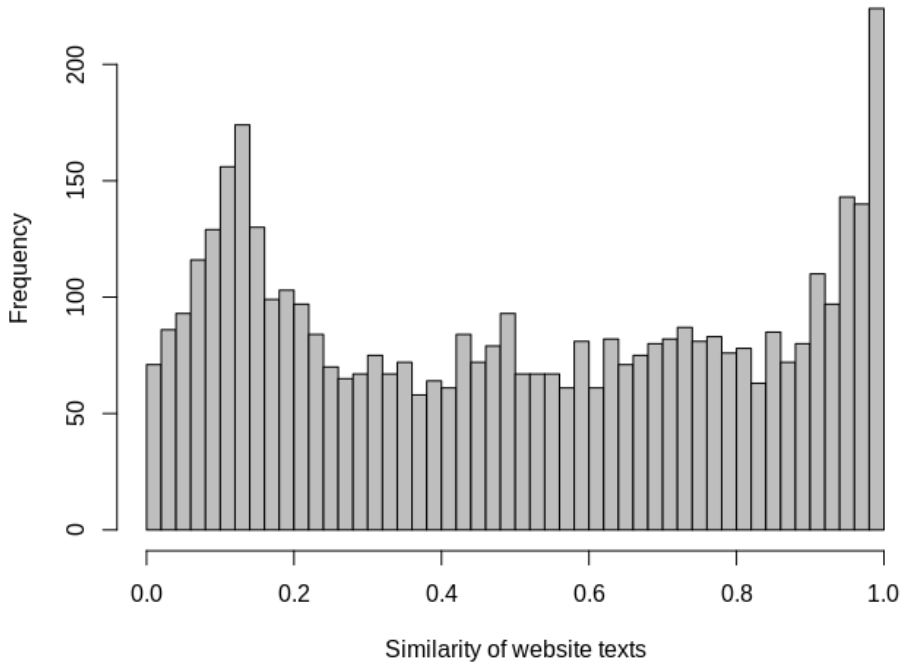


Figure 3. Distribution of the similarity measures of the words extracted from the 4,478 company websites included in the first and last data set. A 1 indicates that all the words on both versions of the website are identical; a 0 indicates no words are in common. The distribution indicates considerable textual changes for many websites.

the combination of both, revealed an average value of 0.50. The distribution of those values indicates a whole range of changes (Figure 3). From this it is obvious that the content of many websites changed considerably over time. To better understand the effect of these changes, the accuracy of the models built on various subsets of the first and last data set and their combinations were determined. The subsets are identified as indicated in Figure 2. By comparing these results, we wanted to know which part(s) of the data contributed to the high accuracy of the first data set. The findings are summarized in Table 1. In this table only the findings of informative combinations are shown.

The results in Table 1 reveal the importance of the data of the units in subset B and the text in subset A for the creation of a highly accurate innovation detection model. In subset B, the website texts of 156 innovative and 129 non-innovative companies are included. Using the data solely in B already results in a model with an accuracy of 76%. Combined with the data in subset A, it is clear that this produces a highly accurate model; an accuracy of 91% is obtained. Subset C on its own has an accuracy of 60%. Adding B to subset C has no additional effect, indicating that the data in C must be of low quality. This becomes obvious when the text in C is replaced by the text of the same units in A (this essentially recreates

subset A). Because of this, the accuracy increases from 60 to 85%. Adding B to the latter increases the accuracy even more; to 89%. Clearly the texts of the first websites are much more informative than the texts in the last dataset. This indicates that -over time- vital information on the innovative, and probably also the non-innovative, character of a company is lost. The effect of B indicates that, in addition, the websites of a set of highly informative units is lost as these websites were no longer active online. The latter suggest that these companies are no longer economically active.

Table 1. Accuracy of the model developed on various combinations of subsets of the first and last scraped data sets to detect technological innovation.

Data set composition	Accuracy (%)	Description
A+B	91	First data set
A	86	Common units, text of first data set
B	76	Unique units in first data set
C+D	60	Last data set
C	60	Common units, text of last data set
D*	-	Unique units in last data set
C+B	60	Common units, text of last data set + unique units in first
C(A)**	85	Common units, text of last data set replaced by text of first
C(A)+B	89	Common units, text of last data set replaced by text of first + unique units in first

*Subset D contained not enough units for model development and did not influence the results of other subsets combinations. **C(A) indicates that the text in subset C is replaced with the text included in A for identical units.

3. Discussion and future work

3.1. Model degradation

From the results described in this paper its obvious that the degradation of the accuracy of the innovation detection model is the result of the combined effect of i) the loss of a part of the company websites included in the CIS survey and ii) changes in the text on the websites that remained active. Both findings are indicative for a dynamically changing population, which is not surprising for the kind of topic studied. It also reveals that innovation is a challenging topic to capture. Because of the high accuracy of the model based on the first data set, the websites of the companies included in the CIS survey must be scraped at a very

appropriate point in time; a time in which the innovative and non-innovative character of the company was clearly reflected in the text on their websites. Future research will focus on comparing the period covered by both survey and web data and will include extracting and studying websites from a web archive. The latter may enable us to determine if and when the optimal point in time occurs between the ending of the CIS survey period and the moment websites should be scraped. What is also apparent from the results is the fact that the findings reveal that it is not the concept that is drifting (changing) but that it is gradually fading away (degrading). Hence, the title of the paper. One of the explanations for this could be a gradual change in the innovative character of the companies studied. Because of this, one even starts to wonder how long the innovation classification data provided by the CIS survey actually holds over time. These are interesting questions that emerge because of the high frequency at which the data can be collected and analyzed.

3.2. Model resurrection

A solution was also developed to deal with model degradation. We were inspired by the literature on dealing with concept drift (Lu *et al.*, 2017) here. The solution most often applied when dealing with this phenomenon is simply retraining the model on new data. In our case this would be the texts derived from 'freshly' scraped websites. From the findings described in this paper its obvious that doing this would never solve the model stability problem; the second data set simply does not contain the information needed for the development of a highly accurate model (see Table 1). We therefore followed the second most often mentioned suggestion in the Machine Learning literature, which is to add newly classified data to the original data set and retrain the model (Janardan, 2017, Gama *et al.,* 2014). We found that this worked well when the classified websites of innovative startups, a total of 855, and those of a large number of websites, i.e. 20,000, for companies in the Business Register were jointly added to the training and test set (Daas & van der Doef, 2020). This increased the training and testset from 4,763 to 25,618 records. Comparing the effect of adding varying amounts of classified Business Register data (samples from 1,000 to 40,000 were tested) on the accuracy of the model, revealed that adding more then 20,000 records did not further improve the model and development took much longer to complete. An accuracy of 88% on the test set was obtained for the new logistic regression model. In our opinion this approach worked well since adding more examples allowed the classifier to find more words which were either positively or negatively related to innovation; i.e. more synonyms were included. This is apparent from the number of words included in the new model; a total of 584. A detailed study of both models revealed that nearly all the words in the old model, 177 to be exact, are included in the new model as well. Future studies will focus on the new model's composition including its stability over time.

References

Aggarwal, C.C. (2016). Mining Text Data. In Aggarwal, C.C. (Ed.) *Data Mining: the Textbook* (pp. 429-455). New York: Springer.

Atefeh, F. & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence,* 31(1), 132-164. doi.org/10.1111/coin.12017.

Daas, P.J.H., Puts, M.J.H. Buelens, B., & van den Hurk, P.A.M. (2015). Big Data and Official Statistics. *Journal of Official Statistics,* 31(2), 249-262. doi.org/10.1515/jos-2015-0016.

Daas, P.J.H. & van der Doef, S. (2020). Detecting Innovative Companies via their Website. *Journal of the IAOS*, accepted for publication.

Florescu, D., Karlberg, M., Reis, F., Rey Del Castillo, P., Skaliotis, M. & Wirthmann, A. (2014). Will 'big data' transform official statistics? Paper for the Quality in Official Statistics Conference, June 2-5, 2014. Vienna. Retrieved from http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf. (Accessed June 2019).

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys,* 46(4), 1-37. doi.org/10.1145/2523813.

Gökk, A., Waterworth, A. & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics* 102(1),653-671. doi.org/10.1007/s11192-014-1434-0.

Höchtl, J., Parycek, P. & Schöllhammer, R. (2015). Big Data in the Policy Cycle: Policy Decision Making in the Digital Era. *Journal of Organizational Computing and Electronic Commerce,* 26(1-2), 147-169. doi.org/10.1080/10919392.2015.1125187.

Janardan, S.M. (2017). Concept drift in Streaming Data Classification: Algorithms, Platforms and Issues. *Procedia Computer Science,* 122, 804-811. doi.org/10.1016/j.procs.2017.11.440.

Jo, T. (2019). *Text mining Concepts, Implementation, and Big Data Challenge*. New York: Springer.

Kim, S-M. & Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp. 1-8. Association for Computational Linguistics. Retrieved from https://www.isi.edu/natural-language/people/hovy/papers/06ACL-WS-opin-topic-holder.pdf.

Kinne, J. & Lenz, D. (2019). Predicting Innovative Firms using Web Mining and Deep Learning. ZEW Discussion paper no 19-001, Mannheim, Germany. doi.org/10.13140/RG.2.2.22526.84809.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS,* 31(3), 471- 481. doi.org/10.3233/SJI-150906.

Lu, J., Liu, A., Dong, F., Gu, F. Gama, J. & Zhang, G. (2018). Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering,* 31(12), 2346-2363. doi.org/10.1109/TKDE.2018.2876857.

Zhang, Y., Chu, G., Li, P., Hu, X. & Wu, X. (2017). Three-layer Concept Drifting Detection in Text Data Streams. *Neurocomputing,* 260, 393-403. doi.org/10.1016/j.neucom.2017.04.047.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11565

A field study on the impacts of implementing concepts and elements of industry 4.0 in the biopharmaceutical sector

Felipe Silva[1], David Resende[2], Marlene Amorim[1], Monique Borges[3]
[1]GOVCOPP/DEGEIT, University of Aveiro, Portugal [2]GOVCOPP/ESTGA, University of Aveiro, Portugal, [3]GOVCOPP, University of Aveiro, Portugal.

Abstract

This article proposes a field study on the applications and impacts of Industry 4.0 (I4.0) in the biopharmaceutical sector, based on an initial literature review. The world is facing a new industrial revolution that is happening at a faster pace than the previous ones. The central idea is the integration between the virtual and the real world through elements that will allow a greater degree of automation and digitization of processes The fieldwork, carried out between July and December 2019, considered semi-structured interviews with managers of pharmaceutical companies or specialists in the I4.0 theme. The interviews pointed out the need for the biopharmaceutical sector to adapt to the concepts of I4.0 and identified its main benefits and barriers. The perceptions were considerably diversified, with the benefits in operational efficiency, productivity and quality being the most scored items. Regarding the main barriers, the most highlighted by the interviewees were the need to break organizational cultural standards, regulatory requirements, the lack of organizational strategies for implementation and the lack of qualified professionals. In conclusion, this work in progress is a contribution to the biopharmaceutical sector and reinforces the imminent need for companies to adapt to this new reality.

Keywords: *Industry 4.0; Pharma 4.0; Biopharma 4.0; Bio 4.0.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

The terms “advanced manufacturing”, “digitization” or “industry 4.0” (I4.0) have been widely cited in the general literature and are directly related to the future of manufacturing and maintaining industrial competitiveness. The world is facing a new industrial revolution that is happening at a faster pace than the previous ones. The central idea is the fusion or integration between the virtual and the real world trough elements that will allow a greater degree of automation and digitization of processes (Figure 1). The term I4.0 originated from a project integrated with the high-tech strategy of the German government, and aimed at the automation and digitization of the manufacturing industry (Daudt & Willcox, 2016).

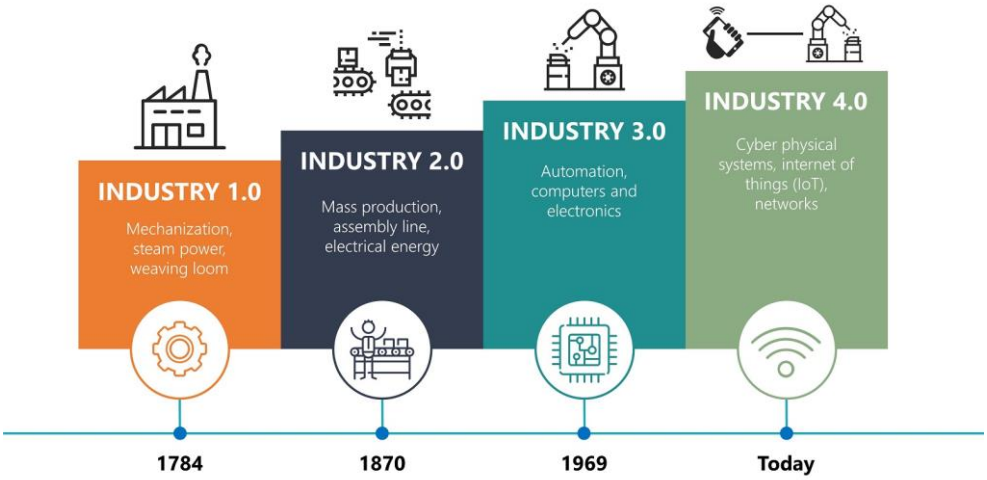


Fig. 1. Industry 4.0 – Fourth Industrial Revolution (Great Myanmar Institute, 2016).

As the concepts of I4.0 are still at an early stage, any attempt to classify their elements becomes complex. However, there is consensus about fundamental elements that underlie the notion ofI4.0, CPS (Cyber-Physical Systems) and the IoT (Internet of Things). The technologies that allow the applicability of I4.0, defined as structuring elements, include automation, Machine to Machine Communication (M2M), Artificial intelligence (AI), BigData analytics, cloud computing, systems integration and cybersecurity. Finally, the complementary elements are those that are not mandatory, but increase the possibilities of application of I4.0, which include 3D printing, RFID tag, QR code, augmented reality and virtual reality, among others (Sacomano & Sátyro, 2018).

1.1. The Biopharmaceutical Industry 4.0

Currently, investments in technological innovation are increasingly migrating from pharmochemical chemicals to biological ones (large molecules produced in living cells). The technological platforms for their production involve the use of living, attenuated or inactivated organisms, whole or in subunits, genetically modified or not, for the production of vaccines or biopharmaceuticals. They require an extraordinary technological challenge, involving expensive and sophisticated activities that include cell culture processes, high performance purification systems, quality control with highly sensitive methodologies, among others (Silva & Caulliraux, 2016).

2. Methodology

The methodology proposed for this study includes the interview of professionals who work in managerial positions in biopharmaceutical industries or I4.0 specialists. In this context, 10 semi-structured interviews were carried out with professionals from different countries between August 2019 and January 2020. The objective was to collect information and perceptions about I4.0 in the biopharmaceutical sector and the expected impacts.

All respondents and the institutions in which they operate had their identities preserved for reasons of industrial secrecy. The interviews were conducted using the semi-structured method, where there is a confluence of previously prepared questions (related to trends, impacts and mapping of the phenomenon studied around the world) with others generated from the responses of the interviewees. The information collected was analyzed by qualitative analysis methods, such as content and narrative analysis.

3. Results

Most respondents perceive a strong need for biopharmaceutical companies to adapt, by implementing the elements the elements of I4.0 in their operations, especially at the risk of losing competitiveness. Some respondents believe in a moderate trend due to industry characteristics and peculiarities that include severe regulatory requirements and risks to people's safety. Factors like these can slow I4.0 applications in the industry.

Table 1. Interviewees in the Field - Biopharmaceutical Sector Managers and Experts at I4.0

Interviewed	Company Characteristics	Company Position	Interviewed Category
A	Portugal Private Pharmaceutical	Industrial Director - Portugal	Biopharmaceutical Manager
B	Portugal Private Pharmaceutical	Digital Transformation Project Manager - Portugal	Biopharmaceutical Manager
C	Transnational Private Bioharmaceutical	Academic Management Program Manager - Spain	Biopharmaceutical Manager
D	Brazilian Private Bioharmaceutical	Production Director - Brazil	Biopharmaceutical Manager
E	Transnational Private Bioharmaceutical	IT Manager - Brazil	Biopharmaceutical Manager
F	Digital Transformation in Production Lines	Digital Latin America - Brazil	Biopharmaceutical Manager
G	Artificial Intelligence (AI) Consulting and I4.0	Head of Machine Learning and Innovation R&D - EUA	I4.0 Specilist
H	Digital Transformation in Production Lines	Director (CEO) - Portugal	I4.0 Specilist
I	Academy Public University – Coimbra	Full Professor of Robotics - Portugal	I4.0 Specilist
J	Private Microconductor Company I4.0 - Portugal	Company Owner and Director - Portugal	I4.0 Specilist

While explaining individual perceptions a diverse set of aspects come to debate. The efficiency gain in operations was the most scored amongst respondents, from the perspective of reducing waste, reducing operating costs and increasing productivity. These benefits are supported by the prospects of real-time quality controls, more detailed process data collection and analysis or predictive equipment maintenance. Quality-related gains, production customization, and reduced time to launch new products were also frequently scored (Figure 3). The literature describes that the adoption of new I4.0-based technologies is expected to improve biopharmaceutical companies towards establishing more robust, flexible and faster manufacturing processes, reducing unplanned outages, fewer equipment defects, better

deadlines, and more agile quality management, analysis and decision making (Markarian, 2018). In addition to the expectation of greater productivity, the interviewees reinforced that real-time monitoring allows the identification of characteristics that make it impossible, for example, to continue a batch, avoiding waste of working hours, energy consumption and raw materials.

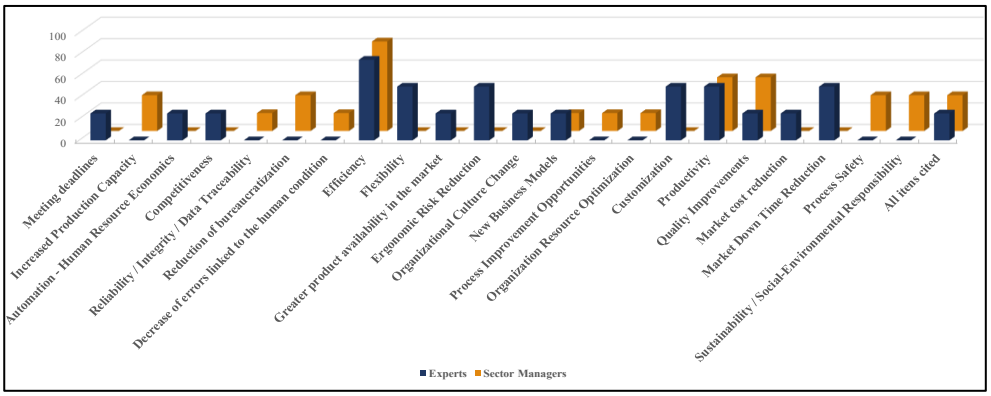


Figure 2. Main benefits identified by respondents of I4.0 in the production of biomedicines.

As described in literature, products obtained in living systems are usually unstable and unpredictable (Silva & Caulliraux, 2016). Therefore, intense and rigorous monitoring in real time is essential. This minimizes the possibility of unwanted products and consequently increases productivity. In a cell fermentation or propagation process, for example, the more closely monitored the reaction medium containing living cells, the greater the possibility of obtaining the desired products. Compared to conventional strategy, the higher level of control and monitoring (based on complex algorithms) will allow analysis and predictive actions to avoid process or equipment failures (Romero-Torres et al., 2018). In this sector equipment is generally aseptic and often subjected to aggressive cleaning agents. Therefore, many of them have significantly high episodes of malfunctions. It is important to detect this predictively so that actions can be taken in advance. Most respondents, especially managers working in the sector, scored this benefit.

Interesting also in this sector is the possibility of increasingly automated production processes, with machines interacting directly with other machines (M2M) without human intervention. In aseptic processes, the presence of any particles (man is the main source) can lead to episodes of batch contamination and damage to millionaires. In the era of I4.0, products are equipped with identifiers, such as QR bar codes or RFID tags, which after being read by the equipment are able to guide the correct sequence of operations through them (Sacomano & Sátyro, 2018) Some respondents antecipate gains in quality and efficiency with the reduction of failures linked to the human condition.

Intelligent processes are expected to be increasingly useful in the biopharmaceutical field. A temperature or pH sensor positioned at a specific point in a bioprocess, for example, can not only transmit the collected values to a computer or cellular exchange, but also compare them to programmed default values and, if necessary, output signals to actuators correct temperature or pH without any human intervention (Sacomano & Sátyro, 2018). The possibility of using intelligent aseptic robots can also be highly interesting in overly repetitive processes or involving physical, chemical or biological hazards (Keller et al., 2018). As reported in the interviews, there is the prospect of extraordinary gains in mitigating ergonomic risks, efficiency, productivity, among others.

Respondents also frequently cited benefits related to digitization and paper replacement. The sensors collect process data and forward it to servers or computers in the cloud for storage and tracking. Punctuated benefits such as socio-environmental responsibility, reliability of data storage under GMP conditions, reduction of bureaucracy and quality are directly related to the theme. The direct connection between CPS, IoT, Big Data, AI and cloud computing will allow the analysis and crossing of this data in a more precise and profound way, resulting in several other benefits mentioned (efficiency, productivity, optimization of resources, among others). Research shows that approximately 70% of the data collected today in the biopharmaceutical industry is unused (Manzano & Langer, 2018).

The possibility of more personalized medicines produced on more flexible production lines was also punctuated in the field interviews. This is an interesting perspective because it is common for certain biomedicines (therapeutic proteins, for example) to serve a small and targeted number of patients. Elements in I4.0 are expected to be able to offer leaner production processes that compensate for production on smaller scales and offer affordable market costs. Factories designed in modular structures, for example, allow greater flexibility and customization in the manufacture of products (Hammer, 2018). The range of single use systems or equipment is another recent challenge in the biopharmaceutical industry. A bioreactor, for example, requires significant manual efforts to connect all the necessary parts to the process. A system can be configured to require the operator to register the barcode on each disposable part to ensure proper component assembly and registration (Markarian, 2018). Virtual or augmented reality tools can also help in this regard, avoiding errors, waste and increasing the security of the process. All of these benefits were raised by the interviewees. Another point cited by the interviewees, especially the experts, is the expectation that more products will be made available to consumers more quickly. Among other elements of I4.0 used for this purpose (Big Data, 3D Printing etc.), modern virtual simulators, for example, quickly manufacture and evaluate products under development, without the need for physical occupation of the factory.

Regarding the main barriers and risks, opinions were quite varied. Reference goes to aspects commonly discussed in the literature, as the difficulty in breaking organizational patterns and

cultures, the stringent regulatory requirements of the biopharmaceutical sector, the lack of definition of strategic alignment for the implementation of the I4.0 elements, high investment and lack of qualified professionals. The interviewees also stressed that it is no use for companies to invest heavily in technological elements of I4.0 if they do not consider strongly the aspects mentioned.

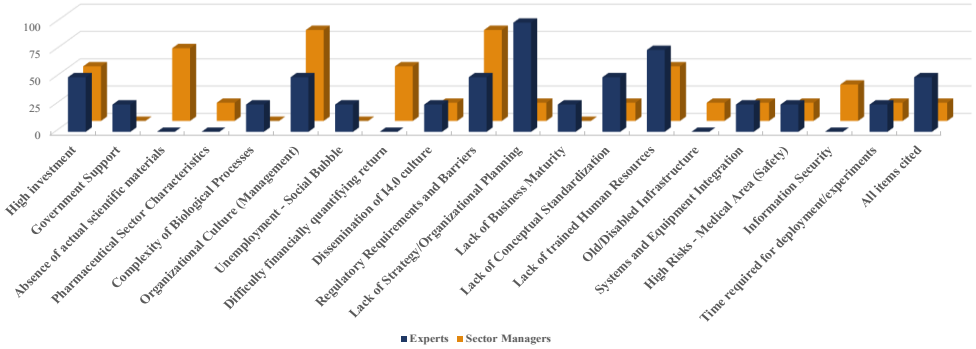


Figure 3. Main barriers identified by respondents in the adoption of I4.0 in the production of biomedicines.

Factors such as low or old IT infrastructure, poor dissemination of I4.0 culture in society (customers, suppliers, regulatory agencies, etc.), lack of business maturity and difficulties in system integration were also cited. The integration between the different systems is extremely challenging point to I4.0. Efforts to standardize the systems are being made among manufacturers to reduce this problem (Romero-Torres et al., 2018). The fragility of information security systems was also addressed in the interviews. It is good practice to separate business data from GMP (Good Manufacturing Practice) production data using a kind of "industrial segregated zone" (Markarian, 2018). Keeping security levels current with the amount of data as a result of increasing connectivity between systems, equipment, and processes that communicate through IoT is highly challenging. The “social bubble” caused by automation and consequent unemployment was also pointed out as a growing concern.

Some studies have been helpful in comparing I4.0's slower pace of insertion into the biopharmaceutical industry than in other areas, such as a Silicon Valley semiconductor company (Romero-Torres et al., 2017). As corroborated by the interviewees, this can be justified by some sector peculiarities. In addition to the regulatory issues inherent in products directly related to life, there is the complexity of biotechnology processes (reported especially by biopharmaceutical managers) and the absence of concrete cases in the literature (little exchange of experience). About the last, most of the documents were found in electronic journals specialized in the biopharmaceutical sector and not in the traditional scientific bases. Probably this fact stems from the scientific incipience of the subject.

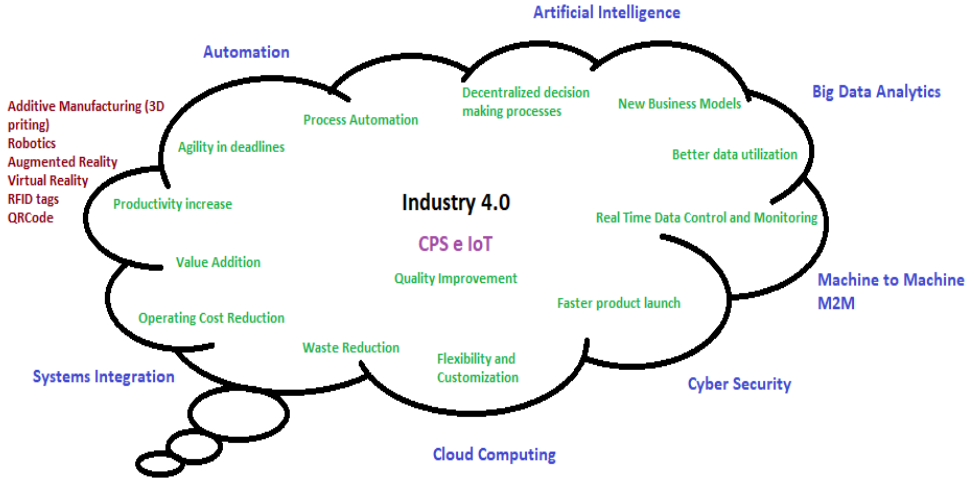


Figure 4. The main technological elements of I4.0.

The figure above summarizes the technological elements of I4.0 and some of the main benefits mentioned in the field interview. In the perception of the interviewees in general, it is essential that companies carefully align the organization's strategies and needs with the available technological elements. It is of utmost importance to carry out a prior assessment on the conditions of integration between systems, equipment and technologies.

4. Conclusion

Overall, field interviews reinforce not only the trend, but the need for biopharmaceutical companies to move into the I4.0 era, especially at the risk of losing competitiveness. The field results, quite diversified, confirmed the main impacts reported in the literature, in addition to bringing others. As for the benefits, efficiency was the main point raised, followed by gains in quality and increased productivity. Gains related to personalization, reduced ergonomic risks and reduced time to market were also scored. Regarding the main barriers, questions were raised that even contribute to a slower pace of implementation of I4.0 than in other sectors. Strict regulatory aspects, the need to break organizational cultures, the lack of qualified professionals, the lack of strategic alignment for implementation and high investment were frequently scored. Others, such as, the high complexity of biological processes and the threat of a "social bubble" due to the reduction in job availability caused by process automation were also addressed. The low number of scientific articles focused on the area still shows incipience and presents the opportunity for new publications, in addition to encouraging organizations to move into the I4.0 era. It is essential to deepen the exchange of experience and knowledge between producers (the entire logistics chain), regulatory

agents and society itself. As a limitation of the study, we can consider the low number of respondents and the consequent impossibility of statistical treatment to provide more consistent conclusions. The expectation is that a new study will be carried out with a more in-depth systematic review, the participation of more respondents and the identification of case by area companies that have already gone through the digitization process.

References

Daudt, G., & Willcox, L. (2016). Reflexões críticas a partir das experiências dos Estados Unidos e da Alemanha em manufatura avançada. In *BNDES Setorial* (44th ed., pp. 5–46). Rio de Janeiro.

Great Myanmar Institute. (2016). Revoluções e o Mundo. Retrieved January 22, 2020, from http://www.greatmyanmarinstitute.com/revolution/

Keller, M., Baum, G., Schweizer, M., Bürger, F., Gommel, U., & Bauernhansl, T. (2018). Optimized Robot Systems for Future Aseptic Personalized Mass Production. *Procedia CIRP*, *72*, 303–309. https://doi.org/10.1016/j.procir.2018.03.066

Manzano, T., & Langer, G. (2018). Getting ready for pharma 4.0 - Data integrity in cloud and big data applications. *International Society for Pharmaceutical Engineering (ISPE)*, 72–19. https://doi.org/10.1177/030857599602000113

Markarian, J. (2018). Industry 4.0 in Biopharmaceutical Manufacturing - Modern technologies offer opportunities to increase manufacturing efficiency. *BioPharm International*, *31*(7), 36–38. Retrieved from http://www.biopharminternational.com/industry-40-biopharmaceutical-manufacturing-0

Romero-Torres, S., Moyne, J., & Kidambi, M. (2017). Towards pharma 4.0; Leveraging lessons and innovation from Silicon valley. *American Pharmaceutical Review*, *7*(1).

Romero-Torres, S., Wolfram, K., Armando, J., Ahmed, S., Ren, J., Shi, C., … Guenard, R. (2018). Biopharmaceutical Process Model Evolution- Enabling Process Knowledge Continuum from an Advanced Process Control Perspective. *American Pharmaceutical Review*. Retrieved from https://www.americanpharmaceuticalreview.com/Featured-Articles/352447-Biopharmaceutical-Process-Model-Evolution-Enabling-Process-Knowledge-Continuum-from-an-Advanced-Process-Control-Perspective/

Sacomano, J., & Sátyro, W. (2018). Indústria 4.0: conceitos e elementos formadores. In Edgard Blucher Ltda. (Ed.), *Indústria 4.0: conceitos e fundamentos* (pp. 27–47). São Paulo.

Silva, F., & Caulliraux, H. (2016). A Desverticalização no Setor de Produção de Biomedicamentos e a Utilização das Empresas CMOs (Contract Manufacturing Organization). *Produto & Produção*, *17*(4), 1–18.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11599

High order PLS path modeling to evaluate well-being merging traditional and big data: A longitudinal study

Francesca De Battisti, Elena Siletti
Department of Economics, Management and Quantitative Methods, Universita' degli Studi di Milano, Italy.

Abstract

We propose using high order partial least squares path modeling (PLS-PM) to define a synthetic Italian well-being index merging traditional data, represented by the Quality of Life index proposed by "Il Sole 24 Ore", and information provided by big data, represented by a Subjective Well-being Index (SWBI) performed extracting moods by Twitter. High order constructs allow to define a more abstract higher-level dimension and its more concrete lower-order sub-dimensions. These layered constructs have gained wide attention in applications of PLS-PM; many contributions in literature proposed their use to build composite indicators. The aim of the paper is to underline some critical issues in the use of these models and to suggest the implementation of a new adapted repeated indicator approach. Furthermore, following some recommendations proposed on the use of PLS-PM in longitudinal studies, we compare the situation in 2016 and 2017.

Keywords: *Well-being; big data;PLS-PM, SEM, hierarchical models.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Several contributions deal with the use of PLS-PM to assess a hierarchical construct model (Tenenhaus et al. 2005). Briefly, in Wold's original design it was expected that each construct would be necessarily connected to a set of observed variables (Wold 1982); on this basis, Lohmöller (1989) proposed the so-called hierarchical component model; recently, Wetzels et al. (2009) provided guidelines outlining four key steps to define a hierarchical construct model, while Becker et al. (2012) focused on the second-order hierarchical latent variable models, which are usually treated with reflective relationships, paying attention to formative relationships; finally, Sarstedt et al. (2019) deepened how to evaluate the results of higher-order constructs in PLS-PM using the repeated indicator and the two-stage approaches.

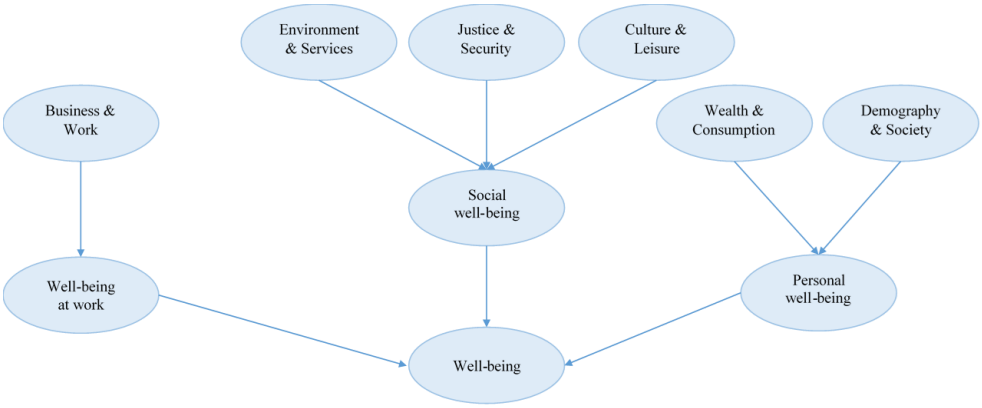


Figure 1. Structural model.

Since 1990, "Il Sole 24 Ore" has provided a yearly quality of life index (QoL) for the Italian provinces (NUTS- 3 in the European nomenclature of territorial units for statistics). Unfortunately, this index is strictly objective, observing only material aspects of quality of life. Following the Stiglitz et al. (2009) recommendation, "current well-being has to do with both economic resources, such as income, and with non-economic aspects of peoples' life (what they do and what they can do, how they feel, and the natural environment they live in)", we suggest to consider an overall index summarizing both the objective and subjective contents, integrating the QoL index with a perceived and subjective source coming from social networks big data: the SWBI index. This last is a multidimensional well-being measure auditing the social networks moods, proposed by Iacus et al. (2015). It is obtained exploiting the big amount of data offered by Twitter data and adopting a new human supervised technique of sentiment analysis (Ceron et al. 2016).

Our proposal consists in mashing-up the different data using the higher order PLS-PM, that has been often applied to buid composite indices using traditional data (Cataldo et al. 2017, Lauro et al. 2018, Davino et al. 2017).

Table 1. : Indicators and themes of "Il Sole 24 Ore" for 2017 and 2016.

Wealth and consumptions	**Demography and society**	**Business and work**
Bank deposits per capita	Resident graduates	Registered enterprises (per 100 inhabitant)
Average monthly rent	Birth rate	Employment rate
Durable goods mean spending for family	Ageing index	Rate of youth unemployment (15-29)
Protests per capita	Internal migratory balance	Loans on deposit (%)
Monthly retirement benefits	Inhabitants for square Km	Exports in % of GDP
Real estate assets per capita (only in 2016)	Acquisition of Italian citizenship	Innovative start-ups per 1000 enterprises
Added value (per capita) (only in 2016)	Number of marriage separations (only in 2016)	Patent application (only in 2016)
GDP per capita (only in 2017)	Average number of education years (only in 2017)	Gender salary gap (%) (only in 2017)
Online shopping (only in 2017)		

Environment and services	**Justice and security**	**Culture and leisure**
Index on urban ecosystem	Home theft	Libraries
Social expenditure of local authorities per capita	Muggings and pick pocketing	Sportiness index
Broadband	Ultra-triennial pending lawsuits	Number of restaurants and pubs
Hospital emigration among regions	Robberies	Foreign traveler expenditure
Number of bank branches, ATM and POS	Scams and computer frauds	Non-profit association
Availability of municipal nursery schools (only in 2016)	Car thefts	Entertainment tickets (only in 2016)
Index of climate changes in temperature (only in 2016)	Index of cause disposal (only in 2016)	Number of cinemas (only in 2016)
Land consumption (only in 2017)	Contentiousness index (only in 2017)	Seats in cinemas (only in 2017)
Spending on drugs (only in 2017)		Number of entertainments (only in 2017)

2. Method

Our data refers to those employed by QoL and SWBI in 2017 and 2016 years. The structural model has been reported in Figure 1.

The definition of the measurement model involves separately three levels. As displayed in Table 1, at the lower order we consider the six themes proposed by "Il Sole 24 Ore", each related to seven indicators.

At the second order of this hieralchical model, we consider the three macro-dimensions suggested by the New Economics Foundation (2012):

- *Personal well-being*
- *Social well-being*
- *Well-being at work*

In defining this measurement level, some issues should be taken into account. First of all, in models with more than two levels the use of the repeated indicator approach yields collinearity problems among constructs, this aspect especially occurs if we need to define them as formative. Furthermore, since the standard structure of QoL, defined in the "Il Sole 24 Ore" project, does not consider a second level, we introduce at this level a new adapted repeated indicator approach, using the related first order indicators and the SWBI components (Table 2). It is worth to notice that, in QoL, some indicators have been changed from one year to the next (Table 1). This is not a good practice in structural equation models, because the invariance of the construct measure has to be ensured. However, our purpose is not to criticize the procedure adopted in the definition of the constructs, but rather to propose a new method to aggregate the constructs and moreover to consider also subjective aspects.

Table 2. : Indicators for macro dimensions.

Personal well-being	**Social well-being**	**Well-being at work**
7 from Wealth and consumptions 7 from Demography and society	7 from Environment and services 7 from Justice and security 7 from Culture and leisure	7 from Business and work
Emotional well-being (SWBI) Satisfying life (SWBI) Vitality (SWBI) Resilience and self-esteem (SWBI) Positive functioning (SWBI)	Trust and belonging (SWBI) Relationships (SWBI)	Quality of job (SWBI)

For the last order, we measure the overall well-being index using all the fifty indicators, applying a traditional repeated indicator approach. All the constructs are defined by composite measures.

Following Davino et al. (2017), to estimate the outer weights of the model we use Mode A for all the higher-order constructs and Mode B for the first-order constructs. In order to estimate inner weights, we use the path weighting scheme.

A drawback in the use of the repeated indicator approach is that the variables with a higher number of corresponding indicators have a major impact on the correspondent higher-order construct. Following this consideration, we foresee that QoL data will impact more on the overall index than data from SWBI. This is not a worrying issue, since the authoritativeness of the QoL is not affected.

2. Results

All the analysis has been carried on using the SEMinR library (Ray et al. 2019).

The PLS-PM path coefficients for the model in 2017 and 2016 are reported in Table 3. The behavior for 2017 and 2016 seams similar. *Personal well-being*, *Social well-being* and *Well-being at work* have a significant effect on the final synthetic well-being index (WB). The *Personal well-being* has the highest impact on WB, nevertheless the number of its indicators is lower than for *Social well-being*.

Analyzing the single macro-dimensions, we notice that on *Personal well-being* the *Wealth and consumptions* has an effect more than twice with respect to *Demography and society* in 2017 and fourfold in 2016. This result is very interesting, considering that in the construction of the "Il Sole 24 Ore" overall index each theme has the same weight. For the *Social well-being*, *Justice and security* has not a significant path in 2016.

Table 3. : Path coefficients for 2016 e 2017 (p-values < 0.001, except for *) and p-value of t-test for equality of paths.

Composite		Path coefficients 2017	Path coefficients 2016	*p*-value
WB	Personal well-being	0.452	0.435	0.245
	Social well-being	0.394	0.402	0.376
	Well-being at work	0.196	0.197	0.464
Personal well-being	Wealth and consumptions	0.720	0.812	0.017
	Demography and society	0.302	0.204	0.016
Social well-being	Culture and leisure	0.338	0.346	0.435
	Environment and services	0.320	0.416	0.039
	Justice and security	0.423	0.307*	
Well-being at work	Business and work	0.999	0.996	0.163

As a measure of goodness of fit of the model, we consider the redundancy (Lohmöller 1989) analyzing the convergent validity, because in hierarchical composite models the R^2 is very closed or equal to 1, as higher-order constructs are almost fully explained by their lower-order constructs. The redundancy index of WB is equal to 0.266 (2017) and 0.292 (2016). The redundancy indices for the sub dimensions are: for *Personal well-being* 0.323 (2017) and 0.337 (2016), for *Social well-being* 0.236 (2017) and 0.262 (2016), and for *Well-being at work* 0.340 (2017) and 0.380 (2016). The average redundancy for the overall model is equal to 0.276 (2017) and 0.300 (2016). These values are not high, but the model is complex and the redundancy values are generally small in PLS-PM; for these reasons, they are judged satisfactory. Furthermore, to validate this model we have to check the collinearity between indicators. The analysis performed at the first order highlights that the VIF values are acceptable, excluding the collinearity issue.

To deep the analysis, we propose to compare the results of 2016 and 2017, following the procedure suggested by Roemer (2016) on the use of PLS-PM in longitudinal studies. Especially , we refer to model type A.1, since our main research object is to investigate the evolution of effects over time and our panel data. After having estimated the model in two different years, we carry on a multigroup analysis, MGA (Henseler et al. 2009), to test the changes in the path coefficients over time; here the different "groups" are interpreted as the different points in time. The last column in Table 3 shows the p-values for the *t*-test used for MGA procedure. The path of *Wealth and consumption* and of *Demography and society* on *Personal well-being* are significantly different in 2016 and 2017. The same happens for the path of *Environment and services* on *Social well-being*

We compare the province rank from the new overall well-being index with the rank from the overall QoL index. The rank correlation indices are: for 2016 *Spearman* = 0.874 and *Kendall* = 0.743, for 2017 *Spearman* = 0.852 and *Kendall* = 0.705. Considering the subjective aspects of well-being slightly affects the classification of the Italian provinces on the basis of the well-being index. However, the different scores obtained are also due to the different methods applied in the aggregation of the dimensions (the same weight for the QoL and different relevance for PLS-PM).

Table 4. : Results of the test of significance of the changes in level of the constructs.

Constructs	Mean difference (2017-2016)	*t*-value	*p*-value
WB	-18656.38	-25.51	0.000
Personal well-being	-20462.18	-24.67	0.000
Social well-being	-759.05	-11.61	0.000
Well-being at work	53.03	57.77	0.000
Wealth and consumptions	-18167.35	-24.34	0.000
Demography and society	-156.01	-15.90	0.000
Culture and leisure	324.73	11.46	0.000
Environment and services	12.25	10.62	0.000
Justice and security	-49.58	-1.64	0.105
Business and work	67.60	58.93	0.000

Finally, being also interested in the change in the level of the constructs over time, we conduct a paired sample *t*-test, performed calculating the non-standardized scores (Table 4). All the constructs have significantly changed during the time. In particular, the WB has got worse from 2016 to 2017, as *Personal* and *Social well-being*. Instead *Well-being at work* has improved.

3. Conclusion

The aim of the paper is to propose a synthetic well-being index using high order PLS-PM merging traditional and Twitter big data. To our knowledge, this is one of the first attempts to merge objective and social network data at provincial level. The novelty of the proposal is also due to the choice of the PLS-PM, with the suggestion of adapting the repeated indicator approach. The insiders could be interested in applying a new approach to take into consideration simultaneously traditional and big data, merging them with suitable weights. The findings are interesting and highlight that considering subjective aspects has an impact on the overall evaluation. Some issues will deserve an in-depth analysis: the estimation of the outer weights in formative-formative hierarchical models and the extension of the multigroup approach to compare more than two situations.

References

Becker J.M., Klein K. & Wetzels M. (2012). Hierachical Latent Variable Models in PLS-SEM: Guidelines for Using Reflective-Formative Type Models. *Long Range Planning*, 45, 359-394.

Cataldo R., Grassia M.G. Lauro N.C., & Marini M. (2017). Developments in Higher-Order PLS-PM for the building of a system of Composite Indicators. *Quality & Quantity*, 51, 657-674.

Ceron A., Curini L. & Iacus S. (2016). iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences*, 105-124.

Davino C., Dolce P. & Taralli S. (2017). Quantile Composite-Based Model: A Recent Advance in PLS-PM, in Latan H. & Noonan R. (eds.) *Partial Least Squares Path Modeling. Basic concepts, methodological issues and applications*, (pp.81-108), DOI 10.1007/978-3-319-54069-3_5, Berlin: Springer.

Henseler J., Ringle C.M. & Sinkovics R.R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing*, 20, 227-319.

Iacus S.M., Porro G., Salini S. & Siletti E. (2015). Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being. ArXiv e-prints 1512.01569.

Lauro N.C., Grassia M.G. & Cataldo R. (2018). Model Based Composite Indicators: New Developments in Partial Least Squares-Path Modeling for the Building of Different Types of Composite Indicators. *Social Indicators Research*, 135, 421-455.

Lohmöller J.B. (1989). *Latent Variable Path Modeling with Partial Lest Squares*. New York: Springer,

New Economics Foundation (2012). *The Happy Planet Index: 2012 Report. A global index of sustainable well-being.*

Ray S., Danks N. & Velasquez Estrada J.M. (2019). SEMinR: Domain-Specific Language for Building PLS Structural Equation Models. R package version 0.7.0.

Roemer E. (2016). A tutorial on the use of PLS path modeling in longitudinal studies. *Industrial Management and Data Systems*, 116 (9), 1901-1921.

Sarsted M., Hair J.F., Cheah J.H., Becker J.M. & Ringle C.M. (2019). How to specify, estimate, and validate higher-order constructs in PLS-SEM. *Australian Marketing Journal*, 27, 197-211.

Stiglitz J.E., Sen A. & Fitoussi J.P. (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress, Tech. Rep. INSEE.

Tenenhaus M., Vinzi V.E., Chaterlin M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis*, 48, 159-205.

Wetzels M., Odekerken-Schröder G. & van Oppen C. (2009). Using PLS Path Modeling for Assessing Hierachical Construct Models: Guidelines and Empirical Illustration. *MIS Quarterly*, 33 (1), 177-195.

Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. Jöreskog & H. Wold (Eds). *Systems under indirect observation*, 2, (pp.1-54). Amsterdam: North-Holland.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11600

Big Data in Corporate Governance decision

Inmaculada Bel, Alfredo Grau
Department of Corporate Finance, University of Valencia, Spain.

Abstract

Progress in Big Data in recent years has grown exponentially, which has allowed the detection and processing of a large amount of data. Until recently, this fact was unattainable by the lack of mechanization of the corporate governance reports. This paper investigates the relationship between corporate governance decisions affect the indebtedness policies of 1,956 industrial companies listed in Europe and the USA over the period 2016–2018 (5,868 observations). To measure corporate governance decisions, we use detailed information on the expertise of audit committees, the proportion of independent directors, board structures and women's presence on corporate boards. Our findings, which are based on a static panel data analysis, show that there is a strong negative relationship between Audit Committees expertise and indebtedness level in European and North American companies. There are also evidence that European and American companies with a one-tier board structure and Audit Committees expertise are less likely to have lower level of indebtedness. Our results shed new light on corporate governance in relation to the experience of audit committees and the influence of their characteristics on indebtedness policy.

Keywords: *Big Data; Corporate Governance; Expertise Audit Committees; Business Analytics.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

One of the consequences of the Internet and global interconnection through the network is the enormous volume of information that organizations and the general public have access to. In recent decades, the challenges and opportunities of Big Data management is a relevant issue in business management in general, and in particular, in financial management. Therefore, it can be analysed the impact of obtaining, managing and analyzing data in the different areas of the company: strategic definition and its implementation, corporate decision making, design of financial policies, etc.

Big Data provides a new vision, a future perspective in order to predict what can happen to take advantage of opportunities and thus, anticipate the events with the use of the techniques provided by the "Business Analytics" area. In this way, you can define analytical models that allow you to model the functioning of organizations. Consequently, it highlights the need for a new paradigm of storage, processing and enhancement of Big Data. Organizations which are move in this philosophy and are generators of information become "Data Driven Business", directed towards decision making as well as strategic management.

In line with above arguments, the objective of this work is to analyze the extent to which corporate governance decisions affect the indebtedness policies of industrial companies listed in Europe and the USA. Particularly, special attention is paid to the effect produced by the previous experience of the Audit Committee in the field of finance on the levels of indebtedness of these companies.

The main findings of this manuscript provide evidence that companies which have a one-tier board structure, have lower levels of indebtedness and if they also have Audit Committees with experience in finance, this reducing effect is softened. These results are generalizable for both Europe and the USA, although this effect is more moderate for North American companies.

2. Theoretical Framework and Hypothesis

Big Data explains extremely large data sets with large storage capacity that generally need to be analyzed using computational methods (Cockcroft and Russell, 2018). In this sense, companies and research centers are deploying a very rigorous computing power to make sense of the huge amounts of data. A large part of the interest that "big" data is that they have a greater potential to contain more interesting patterns and anomalies than "small" data (Cockcroft and Russell, 2018).

Rehman, Chang, Batool and Wah (2016), among others, characterize Big Data for its volume, velocity and value. Subsequently, IBM and Microsoft added one more feature, *veracity*, to describe the reliability of the data. However, and according to Bhimani and Willcocks (2014),

this volume of data which is generated in a continuous and increasing way is largely unstructured. Many of them are likely to be organized in an economically useful sense and quickly processed for decision-making. effective in real time (Krishnan, 2013).

2.1. Impact of Big Data in Finance

The use of Big Data in the financial field has developed in recent years very quickly (Ye and Li, 2017). Despite these advances, there is still little research on how Big Data has influenced the way financial decisions are made, about their impact on strategic responsibilities (Quinn, Dibb, Simkin, Canhoto and Analogbei, 2016), or how this data is handled at the board level (Nutt and Wilson, 2010). However, Big Data offers the potential to reduce risk and improve these strategic decisions by allowing high-level leadership teams to have a more comprehensive vision (Filatotchev and Nakajima, 2010).

Turner, Schroeck and Shockley (2013) consider that Big Data is a source of information and one of the most important assets that organizations have. The financial management business is packed full of transactions that add growing information to the industry. Hence, Big Data offers in finance management the possibility of adopting a more strategic and proactive role within the company (Chua, 2013). In particular, Bhimani and Willcocks (2014) warn against reorienting financial functions to simply harness the potential of big data.

2.2. Hypothesis

In the corporate governance field, Audit Committee (AC) characteristics (such as expertise or independence) are considered relevant factors in order to reduce the opportunistic behaviour of managers and by mitigating agency problems (Madi, Ishak and Manaf, 2014). In this sense, investors demand the presence of audit committees in the companies whose members have relevant expertise (Ghafran and O'Sullivan, 2013). Audit committees with financial expertise are considered an internal monitoring mechanism that can mitigate agency problems and tend to impact on indebtedness policy (Javaid and Javid, 2017). Past research has analysed the effect of some aspects of corporate field with the level of debt such as the independent directors (Doan and Nguyen, 2018), audit committees expertise (Carcello, Hollingsworth, Klein and Neal, 2006), firm size (Harford, Li and Zhao, 2007), Board Structure Type (Calza, Profumo and Tutore, 2017), CEO duality (Harris, 2014), board structure type (Pucheta-Martínez, Gallego-Álvarez and Bel-Oms, 2019), gender diversity (Harris, 2014), among others. The hypotheses to study in Big Data context are:

> ***Hypothesis 1:*** *The greater the financial experience of the AC, the lower the indebtedness.*
>
> ***Hypothesis 2***: *The greater the financial experience of the AC, the greater the negative relationship between the independence board and the indebtedness.*

Hypothesis 3: *The greater the financial experience of the AC, the lower the negative relationship between the board structure type and the indebtedness.*

Hypothesis 4: *The greater the financial experience of the AC, the lower the negative relationship between the women's presence on boards and the indebtedness.*

3. Sample and Variables

The sample used in this study comprised international firm years observations from THOMSON REUTERS EIKON database from 2016 to 2018. This sample included the industrial sector of all the countries belonging to Europe and USA and is grouped in a static data panel with 1,956 industrial companies and 5,868 observations. We have used the industrial sector is due to the fact that this sector plays a very significant role in the global economy.

The series of the variables used (Table 1) have been filtered to eliminate both the observations with errors or absent, as well as those extreme observations in the distributions. This double filtering process has lead to losing approximately 32.8% for the USA and 64.4% for Europe.

Table 1. Description of the explanatory variables.

Parameters	Description
	Leverage Variable
LEV	*Leverage:* Total Debts / Equity
	Main Explanatory Variables
EXA	*Expertise Audit Committees:* Dummy variable that takes the value if the members of the audit committee have financial experience and 0 otherwise.
INDBO	*Independence Board:* Ratio between the proportion of independent directors on boards directors and the total members of the board.
BOTYPE	*Board Structure Type:* Dummy variable that takes the value 1 if the company has a one-tier board structure and 0, if the company has a two-tier board structure.
BGEN	*Board Gender Diversity:* Dummy variable that takes the value 1 if the companies include female directors on corporate boards and 0 otherwise.
	Control Variables
CEODU	*CEO duality:* Dummy variable that takes the value 1 if the CEO of the firm also serves as chairman of the board and 0, otherwise.
LSIZE	*Company Size:* Logarithm of total assets of firms.
ROA	*Profitability:* Profit Before Interest and Taxes / Total Assets.

4. Methodology

In this section, we analyze the determinants of level of indebtedness and will pay special interest to the effect produced by the financial experience of the Audit Committee. We group the large data into a static panel that will allow us, to some extent, to control the unobservable heterogeneity that could occur in the treatment of these data. The econometric approach is:

$$\begin{aligned} \mathrm{LEV_{jt}} = {} & \delta_0 + \delta_1 \mathrm{EXA_{jt}} + \delta_2 \mathrm{INDBO_{jt}} + \delta_3 \mathrm{BOTYPE_{jt}} + \delta_4 \mathrm{BGEN_{jt}} \\ & + (\delta_5 \mathrm{INDBO_{jt}} + \delta_6 \mathrm{BOTYPE_{jt}} + \delta_7 \mathrm{BGEN_{jt}}) * \mathrm{EXA_{jt}} \\ & + \delta_8 \mathrm{CEODU_{jt}} + \delta_9 \mathrm{LSIZE_{jt}} + \delta_{10} \mathrm{ROA_{jt}} + \varepsilon_{\mathrm{jt}} \end{aligned} \quad (1)$$

where LEV_{jt} is the level of indebtedness for industrial sector j in the time period t calculated as the quotient between the total liabilities and equity. δ_0 represents the regression constant. δ_j represents the estimated values of all variables. ε_{jt} are the random perturbations.

The parameters have been estimated by incorporating instrumental variables through the *Generalized Method of Moments* (GMM) to the equation in first differences. To measure the goodness of fit are proposed: adjusted R^2, contrast of Wald set, and estimation error. In addition, the second order serial correlation *m2* test of Arellano and Bond (1991). Furthermore, the over-identification of restrictions Sargan (1958) test. To detect possible multicollinearity problems, we apply the Variance Inflation Factor (VIF). The results obtained for all companies, we confirm the absence of multicollinearity problems since the values of the VIF range between 1.1021 and 7. 7763 (Neter, Wasserman and Kutner, 1989).

5. Results

Table 2 shows the findings for checking all the hypotheses proposed. Moreover, we want to examine the individual effect of independent variables with the indebtedness policy and the moderating effect of audit committee expertise on the other variables.

Table 2. Determinants of indebtedness for industrial firms.

	EUROPE		USA	
Main Variables	**Model 1**	**Model 2**	**Model 1**	**Model 2**
C	-0.173**(-1.971)	0.2912(1.465)	0.537(0.729)	0.594(0.814)
EXA		-0.417**(-2.047)		-0.072*(-0.122)
INDBO		0.035(0.206)		1.152*(0.323)
BOTYPE		-0.108** (-2.727)		-0.068**(-1.366)
BGEN		-0.058(-0.327)		-0.978(-0.304)
Cross Effects				
INDBO*EXA		-0.046(-0.246)		-1.093(-0.288)
BOTYPE*EXA		0.119**(2.418)		0.056*(1.975)
BGEN*EXA		0.194(0.939)		1.08 (0.331)
Control Variables				
CEODU	-0.036(-0.734)	-0.032(-1.162)	0.085*(0.619)	0.015(0.115)
LSIZE	0.042***(9.136)	0.032***(4.966)	0.004**(0.115)	0.001*(0.016)
ROA	-0.953***(-15.543)	-0.934** (5.748)	-0.189**(-1.300)	-0.185**(-1.296)
R^2 *adjusted*	0.0424	0.0828	0.1102	0.1641
Wald	293.99**	9541.85**	2442.76**	12528.60**
Est. error	1.2241	0.9394	1.0378	0.9666
m2 Test	0.92	0.74	0.96	0.88
Sargan Test	62.67(69)	91.66(73)	71.66(69)	76.52(72)

The data correspond to regression results of GMM model in first differences, described in the equation (1). t-Statistic in brackets. Chi-squared: degrees of freedom in brackets for Sargan Test. ***p<0.01, **p<0.05, *p<0.1.

The results obtained for Europe are the following. Model 1 provide evidence that the firm size has a positive (at 1%). However, the profitability presents a negative (at 1%). These results confirm the premise that companies with higher size, results in a higher indebtedness policy and lower levels of return on assets. In Model 2, the coefficient of EXA variable is negative (at 5%). Hence, we confirm the explanatory power of this variable and hence, the, compliance with *Hypothesis 1*. Our evidence suggests that European firms with include this committee tend to support a lower level of leverage, in line with Badolato, Donelson and Ege

(2014). Moreover, we examine the impact of board structure (BOTYPE) and policy of indebtedness. This finding provide evidence that the coefficient is negative (at 5%).

The BOTYPE*EXA is positive and of opposite sign to the main variable in Europe and the USA. This result leads us to accept the *Hypothesis 3*. As a consequence, European firms with a one-level board structure are less likely to have a lower level of indebtedness when there is a greater effect of this committee, in line with Pucheta-Martínez *et al*. (2019).

According to the results for USA, Model 1 provide evidence that the duality of CEO and firm size have a positive (at 10% and 5%, respectively). However, the profitability presents a negative sign (at 10%). These results confirm that when the CEO of the firm also serves as chairman of the board and when the companies have higher size, results in a higher debt policy and lower levels of return on assets.

In Model 2, the coefficient of EXA is negative (at 5%). This finding leads us to accept the *Hypothesis 1*, which suggests that American firms with include an audit committee with directors with financial experience tend to support a lower level of indebtedness. On the other hand, the results also findings that the proportion of independent directors on corporate boards (INDBO) shows a positive sign (at 10%), contrary to our predictions. According to this result, companies which include independent directors tend to increase the indebtedness policy. Furthermore, the variable board structure type (BOTYPE) presents a negative sign (at 5%). Therefore, all companies) with a one-tier board structure and Audit Committees expertise (BOTYPE*EXA) are less likely to have lower level of indebtedness. Furthermore, the variable board structure type exhibits a negative sign (at 10%). Therefore, the *Hypothesis 3* has not to be rejected. Our finding suggests that companies located in USA with a one-tier board structure are less likely to have lower level of indebtedness when there is higher effect of audit committee expertise.

In respect of cross effects analyzed for the proportion of independent directors on corporate boards (INDBO*EXA) and gender diversity (BGEN*EXA), they do not present statistical significance. Consequently, we should reject *Hypothesis 2* and *Hypothesis 4.*

6. Conclusions

The aim of this investigation is to study, in the Big Data environment, the extent to which corporate governance decisions affect the indebtedness policies of industrial companies listed in Europe and the USA. We have paid special attention to the effect produced by the previous experience of the Audit Committee in the field of finance on the indebtedness levels.

The European and North American companies with a one-level board structure are less likely to have a lower level of debt when there is a greater effect of the audit committee's experience. While it is true that the financial formation of this committee, in itself, allows reducing the

volume of debt, when it acts in an organization where the governance structure is unique, this effect is less expansionary. This could be explained by the fact that this financial expertise leads the audit committees to drive an optimal capital structure that does not necessarily imply a simple reduction in indebtedness, but that these levels are the most appropriate for industrial companies listed. Moreover, in board structure unitary when all board members have the same responsibilities and functions, independent directors may be not fulfill their monitory duties. This fact reduces the credibility and objectivity of the board members when monitoring managerial team and ultimately, may reduce in lower level the indebtedness policy. These findings are observed for firms listed in both markets, although it should be noted that this effect has less impact in the USA.

Several implications can be derived from this analysis. Firstly, the findings of this investigation provide evidence that there is a limited presence of female directors on corporate boards. In this sense, our manuscript has a relevant value for government and regulatory bodies, because it allows them to note that there is under-representation of women on boards for Europe and USA, since there is not effect on the leverage with or without crossing effect. Policymakers should recommend the representation of female directors on boards since they behave as a control mechanism that improves the financial decisions of the companies. Second, regulators in Europe and USA should made efforts to consider audit committee members with financial expertise as internal control mechanisms in the companies. This evidence should lead policymakers to consider the benefits to inclusion of financial experts on audit committees to the stakeholder. Third, this evidence may be useful for managers who are willing to enhance the indebtedness policy, as we show that companies reduce the indebtedness if there is Audit Committees expertise and one-tier board structure.

Our study's findings should be considered with caution. The sample used in this study is based on European and North American Companies on the industrial sectors, although the study revealed some factors that are not found in the past research yet. Further research can focus on investigating if the Audit Committee experience has an effect on the making decision process of the indebtedness policy in other countries such as Latin-America or Asia.

References

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and application to employment equations. *Review Economics S.*, 58, 277–97.

Badolato, P., Donelson, D., & Ege, M. (2014). Audit committee financial expertise earnings management: The role of status. *Journal of Accounting and Economics*, 58, 208–230.

Bhimani, A., & Willcocks, L. (2014). Digitisation, "Big Data" and the Transformation of Accounting Information. *Accounting and Business Research*, 44, 469–90.

Calza, F., Profumo, G., & Tutore, I. (2017). Boards of directors and firms' environmental proactivity. *Corporate Governance and Organizational Behavior Review*, 1(1), 52-64.

Carcello, J., Hollingsworth, C., Klein, A., & Neal, T. (2006). Audit committee financial expertise, competing corporate governance mechanisms, and earnings management. Competing Corporate Governance Mechanisms, and Earnings Management.

Chua, F. (2013). *Big Data Essential to Commercial Accountants*.

Cockcroft, S., & Russell, M. (2018). Big Data Opportunities for Accounting and Finance Practice and Research. *Australian Accounting Review*, 28(3), 323–333.

Doan, T., & Nguyen, N. Q. (2018). Boards of directors and firm leverage: Evidence from real estate investment trusts. *Journal of Corporate Finance*, 51, 109-124.

Filatotchev, I., & Nakajima, C. (2010). Internal and external corporate governance: An interface between organization and its environment. *British J. Management*, 21, 591–606.

Ghafran, C., & O'Sullivan, N. (2013). The governance role of audit committees: reviewing a decade of evidence. *International Journal of Management Reviews*, 15(4), 381-407.

Harris, C. K. (2014). Women directors on public company boards: does a critical mass affect leverage?Business and Economics Faculty Publications. 29. Available at: https://digitalcommons.ursinus.edu/bus_econ_fac/29

Harford, J., Li, K., & Zhao, X. (2007). Corporate boards and the leverage and debt maturity choices. SSRN Electronic Journal.

Javaid, H., & Javid, S. (2017). Determining agency theory framework through financial leverage & insider ownership. *International Journal Economics and Finance*, 9, 21-28.

Krishnan, K. (2013). *Introduction to big data. In K. Krishnan* (Ed.), Data warehousing in the age of big data: A volume in MK series on business intelligence. Morgan Kaufmann.

Madi, H., Ishak Z, Manaf, N. (2014) The impact of audit committee characteristics on corporate voluntary disclosure. *Procedia-Social and Behavioral Sciences*, 164, 486–492.

Neter, J., Wasserman, W., & Kutner, M.H. (1989). *Applied regression models*. IL: Irwin.

Nutt, P., & Wilson, D. (2010). *Handbook of decision making*. London: Wiley.

Pucheta-Martínez, M. C., Gallego-Álvarez, I., & Bel-Oms, I. (2019). Board structures, liberal countries, and developed market economies. Do they matter in environmental reporting? An international outlook. *Business Strategy and the Environment*, 28(5), 710–723.

Quinn, L., Dibb, S., Simkin, L., Canhoto, A., & Analogbei, M. (2016). Troubled waters: The transformation of marketing in a digital world. *European J. Marketing*, 50, 2103–33.

Rehman, M., Chang, V., Batool, A., & Wah, T. (2016). Big data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management*, 36(6), 917–928.

Sargan, J. (1958). The estimation of economics relationships using instrumental variables. *Econometrica*, 26, 393–415.

Turner, D., Schroeck, M., & Shockley, R. (2013). *Analytics: The Real–world Use of Big Data in Financial Services*, IBM Global Business Services: 27.

Ye, M. & Li, G. (2017). Big Data, big decisions: Internet big data and capital markets: a literature review. *Financial Innovation*, 3(6), 3–18.

Question-Generating Datasets: Facilitating Data Transformation of Official Statistics for Broad Citizenry Decision-Making

Rahul Yadav[1], Patricia Snell Herzog[2], Davide Bolchini[1]
[1]Indiana University School of Informatics and Computing at IUPUI, United States, [2]Indiana University Lilly Family School of Philanthropy at IUPUI, United States.

Abstract

Citizenry decision-making relies on data for informed actions, and official statistics provide many of the relevant data needed for these decisions. However, the wide, distributed, and diverse datasets available from official statistics remain hard to access, scrutinise and manipulate, especially for non-experts. As a result, the complexities involved in official statistical databases create barriers to broader access to these data, often rendering the data non-actionable or irrelevant for the speed at which decisions are made in social and public life. To address this problem, this paper proposes an approach to automatically generating basic, factual questions from an existing dataset of official statistics. The question generating process, now specifically instantiated for geospatial data, starts from a raw dataset and gradually builds toward formulating and presenting users with examples of questions that the dataset can answer, and for which geographic units. This approach exemplifies a novel paradigm of question-first data rendering, where questions, rather than data tables, are used as a human-centred and relevant access points to explore, manipulate, navigate and cross-link data to support decision making. This approach can automate time-consuming aspects of data transformation and facilitate broader access to data.

Keywords: *Official Statistics; Geospatial Data; Big Data Methods and Automation; Data Economy; Data Access; Data-Based Decision-Making.*

1. Introduction

This paper describes an automation process designed to generate questions from datasets. Questions are key to problem solving, and data-based problem solving is crucial for making informed decisions (e.g. Boss, 2016). Sometimes the power of data is harnessed when a new analysis is run, but often the power of transformation occurs in framing the question. Yet, answers to an unasked question are irrelevant (Gutiérrez, 2013, citing Niebuhr, 1943).

The big data in large administrative datasets can help to answer public policy questions. (e.g. Connelly et al., 2016; Chetty et al., 2014; Chetty, 2013). Harnessing their insights requires understanding the distinctions of the "wide data" arising from few observations relative to variables, often garnered from Internet sources such as Amazon clicks and Twitter likes, from the "long data" with many observations relative to variables, often garnered from administrative databases of official statistics such as tax records and censuses (Chetty 2020). Additionally, asking relevant questions of the big data generated from official statistics has challenges (Connelly et al., 2016; Taylor et al., 2014). Notably, users often need domain-specific knowledge to understand the structure and semantics of these datasets and be able to pose questions their data can answer (Europa, 2017).

Advances in natural query language technologies can automate facets of the question-generating process. For example, Salesforce is using machine learning to facilitate everyday people in querying databases using their natural language (Mannes, 2017). Automating portions of the data access process can aid broader utility of the information embedded in these big datasets for public issues. Despite the many benefits of automating the process of database querying, there are several issues that could impede its development. For one, the domain knowledge embedded within large datasets needs to be treated carefully in the automation construction process in order to facilitate broader accessibility and applicability.

To lessen barriers to data access, this project aims to extract user-relevant meaning from datasets in forms of basic factual questions by automating some of the facets of the data extraction and linking processes. Since geographies are particularly relevant for public policy questions, this project focuses on automating extraction of the geospatial data that are pervasive in datasets of official statistics.

2. Exemplifying Scenario

The following scenario illustrates data access issues surrounding use of official statistics in citizenry decision-making. Residents in a city that lacks funding for adequate public transportation want a new transit line added to alleviate traffic congestion and promote environmental sustainability. Due to the limited infrastructure, there are many opportunities for locations within the city for this new line. Which location is of highest priority?

While transportation services aim to address multiple priorities in the long-term, the limited resources necessitate short-term prioritising of a single new line in a strategic location. In reviewing previous efforts, city officials note that past transit efforts were severely under-funded because many voting citizens declined ballot budget referendums. The resulting lack of funding caused bus delays, driver turnover, and other infrastructure issues. Thus, the goal for the strategic location of the new line is to avoid repeating these previous problems.

To be strategic in selecting the location, city officials decide to target a higher-income area, aiming for wealthier residents to then vote in support of additional funding. The location thus needs to target a location that can facilitate greater commuting to work and cultural activities among high-income residents. To make a data-informed decision about the location that best targets this goal, decision-makers need to know four data points within a set of concurrent geographic units: median household income (higher than city average), median work commute time (higher than within-city average), number of existing public transit lines (lower than city average), and fuel consumption (higher than city average).

In this scenario, the metrics need to be available within-city, at relatively small geographic units that can be mapped along the main roadways. To assess whether the new line is effective, these geospatial data also are needed over time, before and after the new line. Most importantly, decision makers need to be able to access and link relevant data rapidly. While relevant data exist, the wrangling required to extract and merge it is a major barrier.

3. Related Work

Several existing approaches are relevant for these issues. For example, scientists have established that geospatial metadata can be useful for assessing the utility of spatio-temporal data for decision makers (Meeks & Dasgupta, 2004). However, the last decade of attention to how geospatial data contribute to broad citizenry decision-making has centred heavily on the spike in availability of geographic information system (GIS) data generated from user devices (e.g. Bishr & Kuhn, 2007). Indeed, major breakthroughs have occurred in humanitarian assistance as a result of the disaster data that can be rapidly spatially located from widespread use of hand-held devices (Ortiz, 2020). These are important advances.

Yet, advancements from the administrative databases of official statistics have not kept pace. For example, population demographic characteristics, such as median household income levels, remain aggregated within geographic units (GEOIDs: U.S. Census 2018; ANSI: U.S. Census 2019). In these, tracts are a geographic unit that was developed to meaningfully approximate neighbourhoods. Tracts are smaller than metropolitan areas and counties, but larger than city blocks or block groups, and thus tract-level data provide aggregate statistics at within-city geographic units. In the scenario, tract-level data would help to answer questions regarding the most strategic location for the new transit line. One

approach to addressing barriers to broader access to relevant official statistics is to visually represent data through an interactive map (e.g. Cartwright et al., 2013). However, not all relevant data are available within existing maps. For example, Policy Map or Social Explorer are widely used map tools, and SAVI is another tool, which attends to a particular city. In all three of these interactive maps, median household income and work commute time are available from official statistics in the U.S. Census. Yet, number of existing transit routes is only available within SAVI, and average fuel consumption is not available in any.

The U.S. Census uses Federal Information Processing Standards codes (FIPS) to uniquely identify aggregated geospatial data, for example through states, counties, and tracts. In some datasets, FIPS identifiers may be stored as a single string digit, with all three geographic identifiers appended in a single variable. Whereas, in other datasets each geographic unit is stored separately, for example in three variables, each representing state, county, and tract in turn. Moreover, many datasets do not specify the available geographies within the meta-data, requiring domain expertise to discern what geographies are available. The nuances in how aggregated geo-spatial data is identified present challenges for non-experts to access relevant data, and to know at which geo-unit data can be wrangled. While the exemplifying scenario in this paper focuses on U.S. geocodes, problems with non-standardised units also exist at larger geographies, such as nations (Scott & Rajabifard, 2017). Thus, part of facilitating a more accessible data economy of official statistics (Europa, 2017) is to automate the geospatial data transformation process, and thus more readily generate the kinds of questions that can be answered by linked datasets.

4. Question-Generating Datasets

Figure 1 displays the process involved in automatically equipping geospatial datasets with basic questions they are designed to answer. This includes the following six steps.

4.1. Data Sources

To begin an analysis of datasets, we took into consideration different sources of datasets and how they might be stored. Typically based on the type of datasets and its usage, datasets are stored in relational or non-relational databases. For our analysis we had the following requirements for the dataset format: (1) It should be compatible across different platforms; and (2) It should be widely used across industries. Common file formats matched our requirements (Shafranovich, 2005): the dataset is in common tabular format (CSV) and any information available about the dataset is in plain text format (TXT). The data sources that we leveraged were open source datasets available from U.S. Census Data, American Community Survey, Our World in Data, and the Indiana Data Hub, all of which have geographic identifiers available within the dataset.

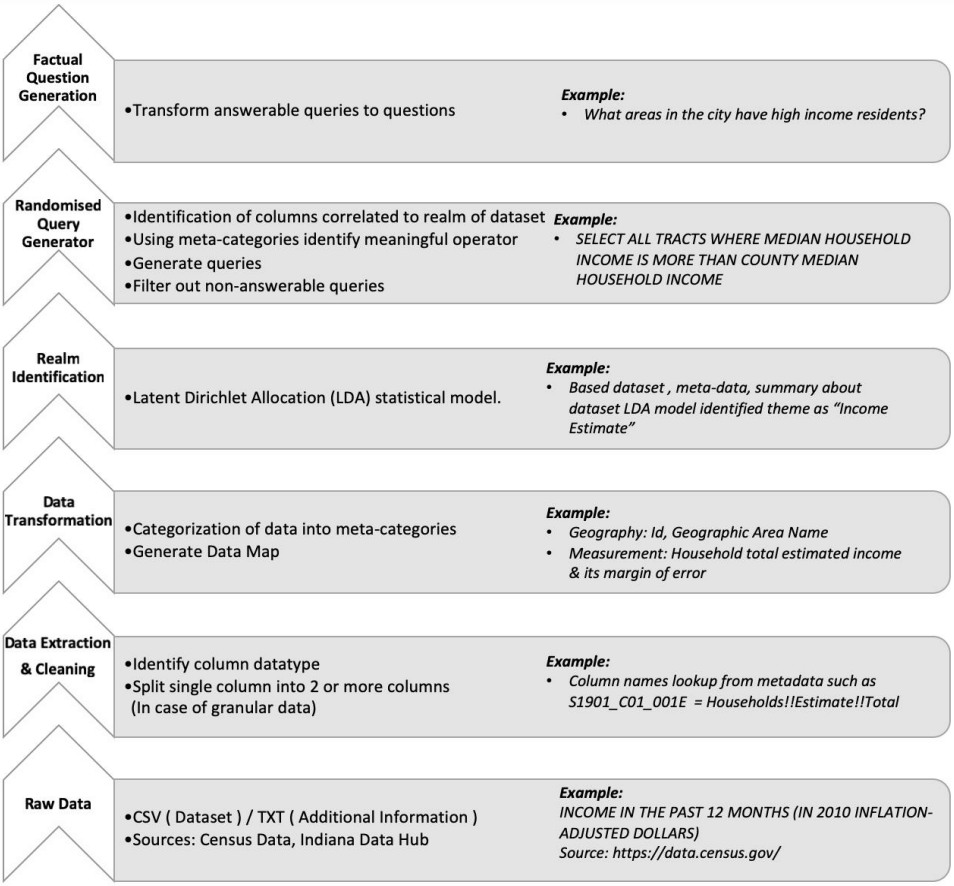


Figure 1. Data transformation process. Source: Author creation.

4.2. Data Extraction and Cleaning

In the process of data extraction (Rahm & Do, 2000), we encountered three issues. First, CSV files do not retain column datatypes. Second, column names often follow patterns that require domain-specific expertise to identify. Third, columns with granular information can be leveraged to gather broader categories. For example, a column with tract FIPS can also contain county and state FIPS within it. Thus, the data extraction process identifies column datatypes and splits granular data into separate variables for state, county, and tract FIPS.

4.3. Data Transformation

Once data are extracted from CSV file, we developed the meta-categories in Table 1, under which most datasets can be represented. We developed a parser (e.g. Srivastava et al., 2017), which leverages column type and segregates datasets into meta-categories based on

patterns and data lookup (e.g. FIPS tables). Once dataset is parsed, the process automatically generates a data dictionary representing the entire dataset in meta-categories.

Table 1. Meta-categories of datasets.

Meta-category	Examples
Time	Date, Period, Duration
Measurement Volume	Length, Weight, Height
Geography	FIPS Code, Latitude, Longitude, Address
Property	Race, Gender, Education

Source: Author creation.

4.4. Identification of Realm

Since we are aiming to generate meaningful answerable questions from the dataset, realm identification is the most important step. With needed meta-data, such as, data summary, data dictionary, and variable description, we were able to correctly identify dataset realms. Realm was processed using a Latent Dirichlet Allocation (LDA) statistical model (Blei et al., 2003) and existing Natural Language Toolkit (NLTK) resources.

4.5. Randomised Query Generator

The next step is to identify columns which are highly correlated to the realm of dataset. Once we select columns which are central to the theme of the data, in conjunction with other randomly selected columns, we leverage meta-category details about these columns to identify meaningful operators for each. Using selected columns and meaningful operators, we generate queries. Once queries are generated, there is a probability that a query might not be answerable by the dataset. Such queries are removed after all queries are generated.

4.6. Factual Question Generation

To render the generated queries as questions, we parsed the queries using semantics (de Marneffe et al., 2006), expressed in Standard Query Language (SQL) into three simple components: Command (action to be taken), Target (specific table) and Additional Clauses (restrictions on the data selection). We replaced the command with question words such as "what" or "which", followed by additional clauses. Examples include:

- Query 1: SELECT RECORDS FROM TABLE WHERE Tract-Level Median Household Income > County-Level Median HH Inc. Q1: ***Which neighbourhoods have high income levels?***

- Query 2: SELECT RECORDS FROM TABLE WHERE Tract-Level # of Public Transit Lines < Median Cty.-Level # of Lines. Q2: ***Which neighbourhoods have low public transit?***
- Query 3: SELECT RECORDS FROM TABLE WHERE Tract-Level Work Commute Time >= Cty.-Level Work Commute Time. Q3: ***Which neighbourhoods have avg.+ work commutes?***
- Query 4: SELECT RECORDS FROM TABLE WHERE Tract-Level Fuel Consumption >= Cty.-Level Fuel Consumption. Q4**:** ***Which neighbourhoods have average+ fuel consumption?***

In this way, datasets can generate examples of the kinds of questions they are equipped to answer, and this automation process can identify those questions in typical semantic terms.

5. Discussion

Traditionally, to interact with datasets, the strategies available to users included: (1) formulating issue specific queries on the data; (2) manipulating and browsing bi-dimensional tables; (3) exploring data through maps and visualisations. Existing paradigms, however, embed a key limitation: they assume that users know or eventually find out which questions to ask the data. However, knowledge workers, who are often not data scientists or domain experts, only vaguely sense the potential value a dataset can hold. As a result, many public datasets are under-utilised. Formulating queries can aid broader access and use.

In this study, we made the first steps towards expanding our understanding of human-data interaction from a data-first to a question-first paradigm. The results of our work exemplify how question-generating datasets can prompt users with examples of potentially relevant questions the dataset can answer. This work sheds light on a potential new class of data-intensive interactive systems, one that endows an available dataset with a suite of available factual questions as the starting point for relevant searches or data exploration.

More broadly, this approach to automating question generation from existing datasets can play an important role in broadening the accessibility and usability of official statistics. The speed at which decisions are made in public and social issues requires that datasets be transformed to aid semantically usable identification of the kind of information a dataset can contribute. In this example, city officials and concerned citizens could more rapidly identify which official statistical databases are equipped to answer their questions, and as a result would be better able to make data-informed decisions regarding where to locate a new public transit line, and its effectiveness in meeting targeted goals over time, for example. Future studies can build upon the automation steps advanced here to improve access for other kinds of data, beyond the geospatial data of this study. Such advancements

would improve widespread data access, increase data-based decision-making for public and social issues, and facilitate informed decisions within the rapid durations necessary.

References

American National Standards Institute (ANSI). (2019, April 23). U.S. Census Bureau. https://www.census.gov/library/reference/code-lists/ansi.html

Bishr, M., & Kuhn, W. (2007). Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In S. I. Fabrikant & M. Wachowicz (Eds.), *The European Information Society: Leading the Way with Geo-information* (pp. 365–387). Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Boss, J. (2016, August 3). The Power of Questions. *Forbes*. Retrieved from: https://www.forbes.com/sites/jeffboss/2016/08/03/the-power-of-questions/

Cartwright, W., Crampton, J., Gartner, G., Miller, S., Mitchell, K., Siekierska, E., & Wood, J. (2013). Geospatial Information Visualization User Interface Issues. *Cartography and Geographic Information Science*, 28(1), 45–60.

Chetty, R. (2013, October 20). Yes, Economics Is a Science. *The New York Times.* https://www.nytimes.com/2013/10/21/opinion/yes-economics-is-a-science.html

Chetty, R. (2020). *Using Big Data to Solve Economic and Social Problems: The Geography of Upward Mobility in America*. Harvard University, Opportunity Insights, Cambridge, MA. Retrieved from: https://opportunityinsights.org/course/

Chetty, R., Hendren, N., Kline, P., Saez, E., & Turner, N. (2014). Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility. *American Economic Review*, 104(5), 141–147.

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, 59, 1–12.

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006, May). Generating Typed Dependency Parses from Phrase Structure Parses. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). LREC 2006, Genoa.

Europa. (2017). Building a European Data Economy. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Retrieved from: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=41205

Mannes, J. (2017, August 29). Salesforce is using AI to democratize SQL so anyone can query databases in natural language. *TechCrunch*. Retrieved from: http://social.techcrunch.com/2017/08/29/salesforce-is-using-ai-to-democratize-sql-so-anyone-can-query-databases-in-natural-language/

Meeks, W. L., & Dasgupta, S. (2004). Geospatial information utility: An estimation of the relevance of geospatial information to users. *Decision Support Systems*, 38(1), 47–63.

Niebuhr, R. (1964). Nature and Destiny of Man, vol. II: Human Destiny (1st Paperback Edition edition). Charles Scribner.

Ortiz, D. (2020). Geographic Information Systems (GIS) in Humanitarian Assistance: A Meta-Analysis. *Pathways: A Journal of Humanistic and Social Inquiry*, 1(2).

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, 23.

Scott, G., & Rajabifard, A. (2017). Sustainable development and geospatial information: A strategic framework for integrating a global policy agenda into national geospatial capabilities. *Geo-Spatial Information Science*, 20(2), 59–76.

Shafranovich, Y. (2005). Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC, 4180, 1-8.

Srivastava, S., Labutov, I., & Mitchell, T. (2017). Joint concept learning and semantic parsing from natural language explanations. Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, 1527–1536.

Taylor, L., Schroeder, R., & Meyer, E. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? Big Data & Society, 1(2), 1-10.

Understanding Geographic Identifiers (GEOIDs). (2018, October 10). U.S. Census Bureau. https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11609

Evaluating accredited mHealth applications. An exploratory study

Ana Gessa, Amor Jiménez, Pilar Sancha
Department of Financial Economics, Accounting and Operations Management, University of Huelva, Spain.

Abstract

Standards and validation practices regarding mobile health apps need to be established to ensure their proper use and integration into medical practice.

This preliminary study aims to conduct a comparative analysis of the entire apps that have been awarded by a Quality Seal to identify significant differences according to the variables analyzed (user, developer, category and consumer ratings) and identified quality attributes.

Although the applications analysed are characterised by their heterogeneity, this research found that seven out of 50 remarkable attributes had significant influence on the application evaluation process, according to the recommendations on design, use and assessment of health from AppSaludable. Only some attributes (adaptation of contents, pilot testing, and accessibility and usability) were correlated with some apps' features.

This study can contribute to improving both the processes of validation and quality of medical care of the citizens and in general, the medical practice.

Keywords: *health app; Appsaludable; quality seal; eHealth; telecare.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

Digitalisation and technology are present throughout our society and influence all areas of our lives. This development has extended to health care and medical decision-making. The latest advances in mobile communications and technologies have led to the implementation of electronic health records and a huge spectrum of portable wireless devices (mobile health; m-Health) which people can use to transmit, store, process and retrieve real-time and non-real-time data between patients and medical personnel or between medical personnel (Hansen, Sanchez-Ferro & Maetzler, 2018; Adibi, 2012).The progressive increase of health apps (h-apps) has led to debate or discussion in recent years to ensure that mobile technology can have a huge impact on healthcare quality and citizens' health. Because the recipients of these apps can be healthcare professionals, medical and nursing students, patients and the general public, the devices can be a valuable tool in health care management (Mosa, Yoo & Sheets, 2012; Paglialonga, Lugo & Santoro, 2008).

It is necessary to ensure the quality of these applications, the identified main problems that a health app must address in order to be of quality are security; data protection and reuse of data; possible risks related to misuse; poor regulation; and lack of standards for validation, efficiency and quality (BinDhim, Hawkey & Trevena, 2015; Huckvale, Prieto, Tilney, Benghozi & Car, 2015; Martínez-Pérez, De La Torre-Díez & López-Coronado, 2014).

Medical applications must be regulated and thoroughly reviewed, and carry out a series of measures to improve the development of evidence-based medical applications while maintaining their open character (Buijink, Visser & Marshall, 2013). There are studies whose aim was to propose a simple, objective, and reliable tool for classifying mobile health apps with a set of criteria and assessing their quality (McMillan, Hickey, Patel & Mitchell, 2016; BinDhim et al. 2015; Stoyanov, S. R., Hides, L., Kavanagh, D. J., Zelenko, O., Tjondronegoro, D., & Mani, M. 2015).

In line with all this, there are numerous works that identify the need, on one hand, for a self-certification model for medical apps (Lewis, 2013) and, on the other hand, to analyse the variables that make them unsafe and of poor quality.

Despite the benefits they offer, better standards and validation practices regarding mobile medical apps need to be established to ensure the proper use and integration of these increasingly sophisticated tools into medical practice. This will help to improve the existing tools and may lead to a better comprehensive m-Health app assessment tool (Nouri, R., Niakan Kalhori, S., Ghazisaeedi, M., Marchand, G., & Yasini, M. 2018).

Aware of the importance of h-apps in the medical practice, in this paper, we provide an comparative analysis of h-apps with remarkable quality and safety. They have been awarded the AppSaludable Quality Seal. Specific objectives were to: a) Identify the remarkable

attributes of these apps—both the favourable and those that need improvement—and; b) assess significant relationships among the analysed attributes and selected variables.

The reflections derived from this work can help improve the processes of accreditation of the quality and safety of health applications. These are necessary mechanisms to correct the distortions or bad practices generated by the informational asymmetry inherent in the mobile environment.

2. Material and Method

Mixed methods research was selected to achieve the objective of this work. The procedure followed is shown in Figure 1.

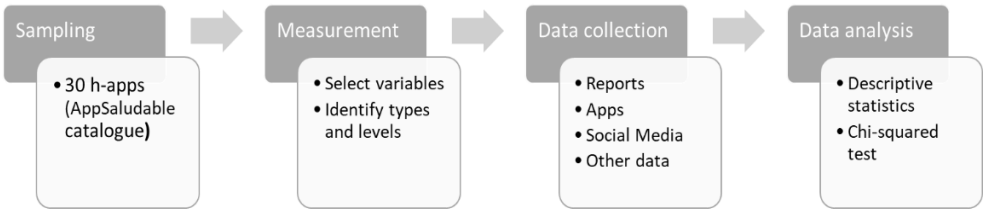


Figure 1. Research methodology.

According to AppSaludable Quality Seal Requeriments (Andalusian Agency for Healthcare Quality, 2012) at the end of April 2019 (http://www.calidadappsalud.com/), 30 health applications were certified. We have analysed all the certified apps. This seal is the first Spanish seal that recognises the quality and safety of health apps. It is free and open to all public or private apps, both Spanish and from other countries. It is a guarantee seal used in order to recognise reliable mobile apps.

This seal is based on 31 recommendations, grouped into four broad categories containing 16 criteria and published in the Guide of Recommendations on Design, Use and Assessment of Health Apps (see Figure 2).

For the analysis of the apps, we collected descriptive information on each app (e.g. price, platform, certification date, etc.) as well as on its technical aspects (e.g. developer, sharing capabilities, etc.). Additional sections collect information on the target user group, as well as information on aspects of the app of interest for the study. These domains may be adapted to include/exclude specific content areas as needed. Next, for each app we extracted the number of downloads and average user rating score. Two variables were identified to evaluate the app: (1) the most frequent remarkable aspects and (2) the issues that need improvement.

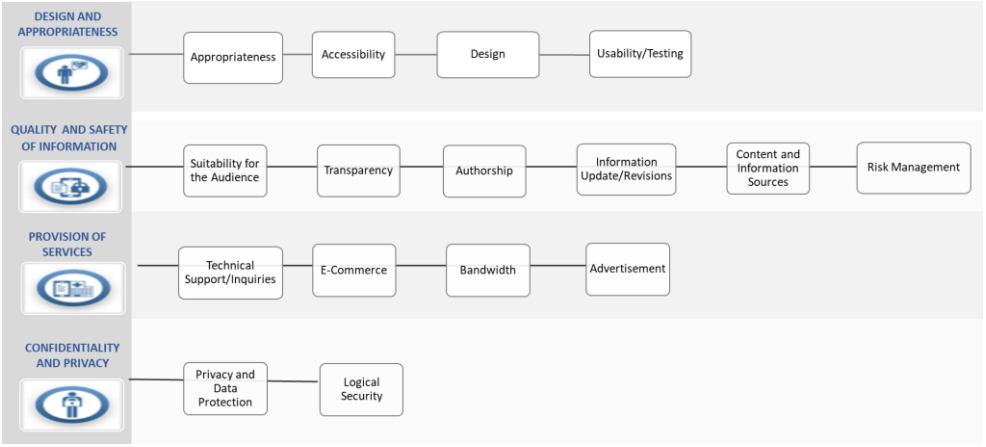


Figure 2. Recommendations on design, use and assessment of health Apps.
Source: http://www.calidadappsalud.com/en/listado-completo-recomendaciones-app-salud/

Data for these variables were taken directly from the Health Apps Catalogue Reports resulting from the evaluation of conformity assessment body. The rest of the variables were obtained by analysing the applications and other specific websites, rankings or social networking, etc. Descriptive statistics were used to summarise h-apps' key characteristics. Some variables, such as platforms, users, main categories and developers, were analysed. In addition, the most frequent remarkable attributes and issues needing attention have been identified using Pareto distribution. Possible relationships between variables were studied through hypothesis test with the Chi square statistic. Analysis results are shown below.

3. Results and Discussion

3.1. Descriptive statistics

Table 1 summarizes the main characteristics for the 30 mobile health-apps.

Table 1. Details of h-apps (n=30)

Characteristics	Frequency		Characteristics	Frequency	
	No.	%		No.	%
Users			*Seal date*		
Patients	10	33.34	Before 2016	15	50.00
Professionals	13	43.33	2016	2	6.67
General public	7	23.33	2017	2	6.67
Category			2018	9	30.00
Monofunction	21	70.00	2019	2	6.66
Multifunction	9	30.00	*Interaction*		
Developer			Yes	5	16.7
Health sector company	6	20.00	No	25	83.3
Technological company	10	33.3			
Sanitary professional	10	33.3			
Particular	1	3.4			
Public administration	3	10.00			

3.2. Statistical analysis

This research shows that seven out of 50 remarkable attributes had significant influence on the application evaluation process according to the AppSaludable recommendations on design, use and assessment of health. The three attributes that are valued the highest are adaptation of the contents to the audience (QA 1), pilot testing (QA 2) and application of universal design principles (QA 9). The four main improvement proposals are editorial commitment to review contents (IP 11), scientific-technical quality control (IP 1), improvement of the accessibility and usability (IP 2) and content selection criteria (IP 18). Results are shown in Figure 3.

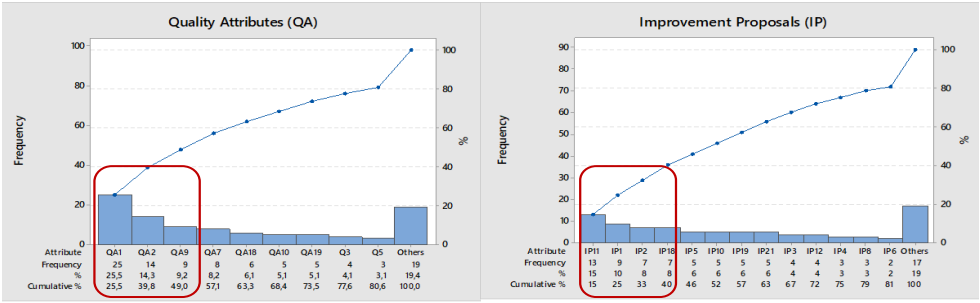


Figure 3. Most frequent quality attributes and improvement proposals of the h-apps.

In order to assess significant relationships among the attributes analysed, a chi-squared test was conducted for the 30 h-apps. Table 2 summarises association test results between the most positive attributes of the h-apps and different selected variables.

Table 2. Chi-squared test for independence between attributes (QA) and the type of user, category, and developer.

	User		Category		Developer	
Attributes	**Chi-squared**	**p**	**Chi-squared**	**p**	**Chi-squared**	**p**
QA 1	0.733	0.693	3.981	0.046	1.298	0.9
QA 2	3.548	0.170	0.408	0.523	9.128	0.05
QA 9	3.527	0.171	0.068	0.794	3.958	0.4

Results do not show significant differences in the evaluation of the apps attributable to the type of user. With regards to the category variable, the association is statistically significant ($P < \alpha = 0.05$) with attribute QA 1; it depends on whether the app has only one function (health and general welfare, medical information, remote monitoring and sensor-based or other) (64%) or whether it has more than one function (36%).

In the case of the type of developer, the association is statistically significant ($P < \alpha = 0.05$) with attribute QA 2. The realisation of pilot tests is present in apps created by companies and professionals from the health sector. These represent 65% of the apps evaluated.

For the issues which need attention, Table 3 shows the results of our statistical analysis.

Table 3. Chi-squared test for independence between improvements proposals (IP) and the type of user, category, and developer.

	User		Category		Developer	
Attributes	**Chi-squared**	**p**	**Chi-squared**	**p**	**Chi-squared**	**p**
IP 11	3.234	0.199	0.068	0.794	5.545	0.22
IP 1	0.784	0.676	1.190	0.275	3.721	0.45
IP 2	0.088	0.957	0.524	0.469	9.950	0.045
IP 18	1.503	0.472	1.074	0.3	4.629	0.35

A significant relation was found only between accessibility and usability, and the developer. This calls attention to how the developed apps by technology companies and healthcare personnel represent 77% of the evaluated apps that need to improve that attribute.

Finally, analysis results do not show a relationship between the user experience and 'average user rating score'. However, the apps with the highest user ratings (score > 4, on a scale of

1-*strongly disagree*- to 5-*strongly agree*-) comprise the highest percentage for the highlight attributes (86.6%, 76.9% and 85.7% for attributes QA 1, QA 2 and QA 9 respectively) (see Table 4).

Table 4. Chi-squared test for independence between attributes (QA/IP) and the users' opinion.

		Users opinion (scale 1-5)					
Attributes		**≤ 2**	**2 < x ≤4**	**>4**	**Total**	**Chi-squared**	**p**
QA 1	**Yes**	0	3	19	22	0.131	0.93
	No	0	1	4	5		
QA 2	**Yes**	0	3	10	13	1.356	0.244
	No	0	1	13	14		
QA 9	**Yes**	0	1	6	7	0.002	0.963
	No	0	3	17	20		
IP 11	**Yes**	0	2	2	4	0.934	0.334
	No	0	6	17	23		
IP 1	**Yes**	0	1	5	6	0.021	0.99
	No	0	3	18	21		
IP 2	**Yes**	0	2	10	12	0.059	0.809
	No	0	2	13	15		
IP 18	**Yes**	0	1	5	6	0.021	0.99
	No	0	3	18	21		

4. Conclusions

There is no doubt that m-Health is a key factor in the challenge of moving towards more sustainable health, improving efficiency and effectiveness, reducing costs and meeting the main needs of our society. Therefore, in this context, the research reveals the suitability of an assessment tool with a wide scope of application, identifying the most frequent positive and negative h-apps attributes.

The results of this work show that the certification model is in the growth state. Administration is concerned about the correct use of these applications in evolving toward a state of maturity. However, this will only be possible if the weaknesses detected in this study are corrected. In conclusion, we would emphasise the following issues:

- It should be noted the lack of integration in a larger project (national, European or international) that supplies a wide coverage and reliability of the seal application criteria.
- This research has detected that the seal must demand a higher level in technical qualities, content and security. We highlight the recommendations concerning accessibility and usability requirements as pointed out in the 44% of the samples analysed. In particular, this attribute performed significantly with the variable "developer" (see Table 3).
- The wide heterogeneity in assessment criteria for m-Health requires a redefinition of the meanings of each criterion.
- Because of the lack of a significant relationship between remarkable attributes and average user rating score, perhaps a re-examination of the seal evaluation process would be useful.

These initiatives definitely are welcome in the field of medicine, as demonstrated by the 80 apps waiting to obtain the AppSaludable seal. Obviously, it can contribute to improving both the processes of validation and the quality of citizens' medical care.

References

Adibi, S. (2012). Link technologies and BlackBerry mobile health (mHealth) solutions: a review. *IEEE Transactions on Information Technology in Biomedicine*, 16(4), 586-597. http://doi.org/10.1109/TITB.2012.2191295

BinDhim, N. F., Hawkey, A., & Trevena, L. (2015). A systematic review of quality assessment methods for smartphone health apps. *Telemedicine and e-Health*, *21*(2), 97-104. http://dx.doi.org/10.1136/bmjinnov-2014-000019

BinDhim, N. F., & Trevena, L. (2015). Health-related smartphone apps: regulations, safety, privacy and quality. *BMJ Innovations*, *1*(2), 43-45. https://doi.org/10.1089/tmj.2014.0088

Buijink, A. W. G., Visser, B. J., & Marshall, L. (2013). Medical apps for smartphones: lack of evidence undermines quality and safety. *Evidence-based medicine*, *18*(3), 90. http://dx.doi.org/10.1136/eb-2012-100885

Hansen, C., Sanchez-Ferro, A., & Maetzler, W. (2018). How mobile health technology and electronic health records will change care of patients with Parkinson's disease. *Journal of Parkinson's disease*, 8(s1), S41-S45. http://doi.org/10.3233/JPD-181498

Huckvale, K., Prieto, J. T., Tilney, M., Benghozi, P. J., & Car, J. (2015). Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment. *BMC medicine*, *13*(1), 214. https://doi.org/10.1186/s12916-015-0444-y

Lewis, T. L. (2013). A systematic self-certification model for mobile medical apps. *J Med Internet Res*, *15*(4), e89. https://doi.org/10.2196/jmir.2446

Martínez-Pérez, B., De La Torre-Díez, I., & López-Coronado, M. (2015). Privacy and security in mobile health apps: a review and recommendations. *Journal of medical systems*, *39*(1), 181. https://doi.org/10.1007/s10916-014-0181-3

McMillan, B., Hickey, E., Patel, M. G., & Mitchell, C. (2016). Quality assessment of a sample of mobile app-based health behavior change interventions using a tool based on the National Institute of Health and Care Excellence behavior change guidance. *Patient education and counseling*, *99*(3), 429-435. https://doi.org/10.1016/j.pec.2015.10.023

Mosa, A. S. M., Yoo, I., & Sheets, L. (2012). A systematic review of healthcare applications for smartphones. *BMC medical informatics and decision making*, *12*(1), 67. https://doi.org/10.1186/1472-6947-12-67

Nouri, R., Niakan Kalhori, S., Ghazisaeedi, M., Marchand, G., & Yasini, M. (2018). Criteria for assessing the quality of mHealth apps: a systematic review. *Journal of the American Medical Informatics Association*, *25*(8), 1089-1098.

https://doi.org/10.1093/jamia/ocy050

Paglialonga, A., Lugo, A., & Santoro, E. (2018). An overview on the emerging area of identification, characterization, and assessment of health apps. *Journal of biomedical informatics*, *83*, 97-102. http://doi.org/10.1016/j.jbi.2018.05.017

Stoyanov, S. R., Hides, L., Kavanagh, D. J., Zelenko, O., Tjondronegoro, D., & Mani, M. (2015). Mobile app rating scale: a new tool for assessing the quality of health mobile apps. *JMIR mHealth and uHealth*, *3*(1), e27. https://doi.org/10.2196/mhealth.3422

Strategic Open Innovation model:mapping Iberdrola network

Naiara Pikatza, Izaskun Alvarez-Meaza, Rosa María Río-Bélver, Ernesto Cilleruelo
Industrial Organization and Management Engineering Department, University of the Basque Country, Spain.

Abstract

Companies are increasingly obliged to collaborate with each other if they want to be innovative, and this growing transfer of knowledge takes place in a context of Open Innovation. To study these scientific-technological collaboration networks within an Open Innovation context, the case study of Iberdrola, a Spanish business group dedicated to the production, distribution and marketing of energy, has been chosen. Two methods have been used; the bibliometric method to analyze the Iberdrola scientific network, and patent data analysis, to analyze the technological network. This has been achieved by using the Scopus and PatSeer databases, and refining the data with VantagePoint software. It was found that in both cases collaboration takes place with the university, other companies, and research centers. Iberdrola has an extensive scientific and technological collaboration network throughout the world. Both scientific and technological collaboration, despite it not being common practice in companies, nevertheless, it can be concluded that Iberdrola is committed to such collaboration in following with the guidelines of its organizational model based on Open Innovation.

Keywords: *Open Innovation; bibliometric analysis; patent analysis; network analysis.*

1. Introduction

Companies cannot claim to possess all the necessary knowledge needed to develop their innovation work and, therefore, are obliged to collaborate with each other if they want to survive. (Bogers, 2011; Chesbrough, 2003). In this context, Open Innovation has appeared in the last thirteen years. The expression "open innovation" is used to show how organizations work together for innovation praxis, notably the relevance and merit of running knowledge inflows and outflows (Anderson & Hardwick, 2017). Due to the increasing complexity of knowledge, more and more alliances and collaborations are formed between companies, universities and research centers to achieve a scope (Bogers, 2011). In this context, Iberdrola, a spanish business group dedicated to the production, distribution and marketing of energy, is a company where Open innovation is contemplated within its organizational philosophy (Tejedor-Escobar & Martínez-Cid, 2009). Iberdrola promotes the launch of R&D projects in different areas, such as smart grids, alternative and renewable energies and ensuring universal access to energy services, among other things. The company practices an open R&D model and collaborates with universities, technology centers and institutions through programs and agreements to leverage complementary assests and capabilities, and to accelerate the commercialization of innovation (Iberdrola, 2020).

Unlike a conventional literature review, the bibliometric method provides an innovative quantitative process, having been widely used in scientific research as an analytical tool to help academics understand the behavior of science within a given field of research; also known as scientometrics, it allows us to capture and map scientific knowledge (Leydesdorff, 2001; Persson, Glänzel, & Danell, 2004; Zemigala, 2019).

According to the OECD (Organization for Economic Co-operation and Development, 2009), patents provide a uniquely detailed source of information of inventive activity, which is why patents are important for evaluating the performance of industry research and development (R&D) (Griliches, 1990). Hence, patent data analysis makes it possible to capture, analyze and map quantitatively technological developments carried out by an organization or related to a particular technology.

Therefore, the aim of this article is to identify the scientific-technological collaboration pattern of Iberdrola and to examine this collaborative network in a context of Open Innovation, in order to contrast the organizational philosophy established in the Iberdrola company and its collaborative reality.

2. Methodology

The research process is adapted from Bildosola et al. (Bildosola, Río-Bélver, Garechana, & Cilleruelo, 2017) and is based on 3 steps intended to define and analyze the scientific and

technological collaborations of the Iberdrola company. Figure 1 shows the research approach, revealing that each step has its input and output, creating a flow of information that allows the set objectives to be achieved. In each step the specific technique used is identified. However, the objective of this research process is that it can be applied to other case studies.

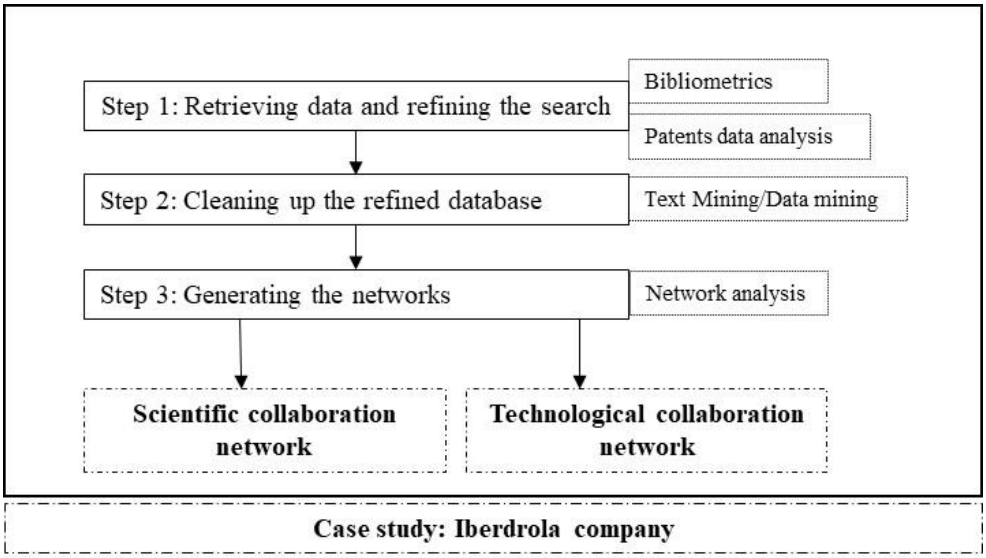


Figure 1. Research process step by step.

Step 1. *Retrieving data and refining the search*. The first task is to generate two specific databases concerning scientific publications and patents related to the Iberdrola company. The specific databases were generated from the Scopus scientific database and PatSeer as the patents database. Scopus is one of largest abstract and citation databases of peer-reviewed literature (75 million documents indexed) (Elsevier, 2017) and it was selected to provide scientific publications. Therefore, the query was built using "IBERDROLA" as affiliation, retrieving a total of 450 documents. The patent analysis was carried out using PatSeer, a complete global patent database and research platform containing the world's most comprehensive full-text Patent collection (Sinha, M.; Pandurangi, 2016). The search for patents has been carried out on the basis of Assignee, obtaining 131 patent families related to the Iberdrola or Iberduero company.

Step 2: *Cleaning up the refined database*. This second task involves the use of text mining tools. The scientific database and the patents database were imported into VantagePoint® (VP) software, text mining software that helps us identify the fields from raw data and show results through a combination of statistics.

Step 3: *Generating the collaboration network*. The networks are divided in two parts: the science collaboration network and the technological collaboration network. The networks are generated and visualized through Gephi software (Bastian, Heymann, & Jacomy, 2009).

3. Results

3.1. Scientific collaboration network

The evolution of publications of Iberdrola is represented in Figure 2. Since 1992 it has collaborated with 30 countries all over the world, and produced 450 articles. However, although the data are not constant, there is a clear upward trend, especially from 2008 (20), with peaks in 2009 (33), 2010 (31), and 2016 (31). Despite the number of publications decreasing substantially (11) in 2019, between January and February 2020 there have already been 14 publications, a clear upward trend.

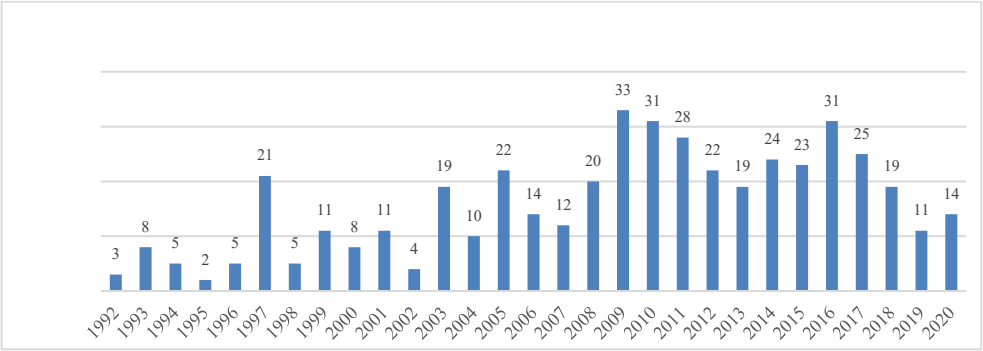


Figure 2. Evolution of publications of Iberdrola.

Regarding institutions, among Iberdola's 10 main scientific collaborators are universities, research centers and other companies (see Figure 3). It should be noted that most of them are Spanish Universities. The Comillas Pontifical University should be highlighted, with more than 50 documents.

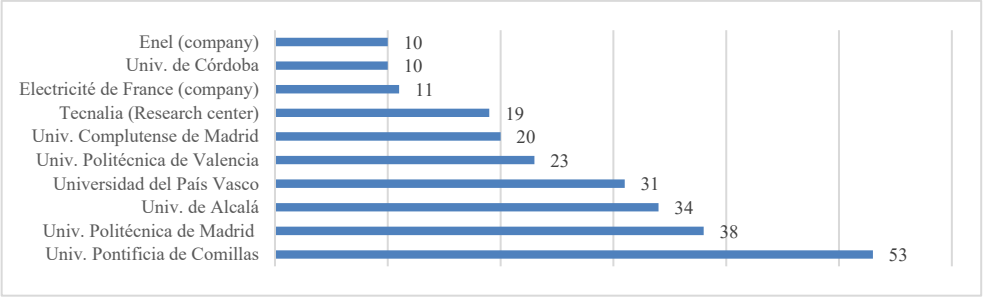


Figure 3: Top scientific institutions

Analyzing the countries collaboration network (see Figure 4), it was found that the United Kingdom (32) has the strongest co-authorship relation with Iberdrola, followed by the USA (28) France (27), Germany (26) and Italy (26), however, it also collaborates with 24 other countries. Iberdrola's commitment to the collaboration of science is clear, with an obvious commitment to internationalization.

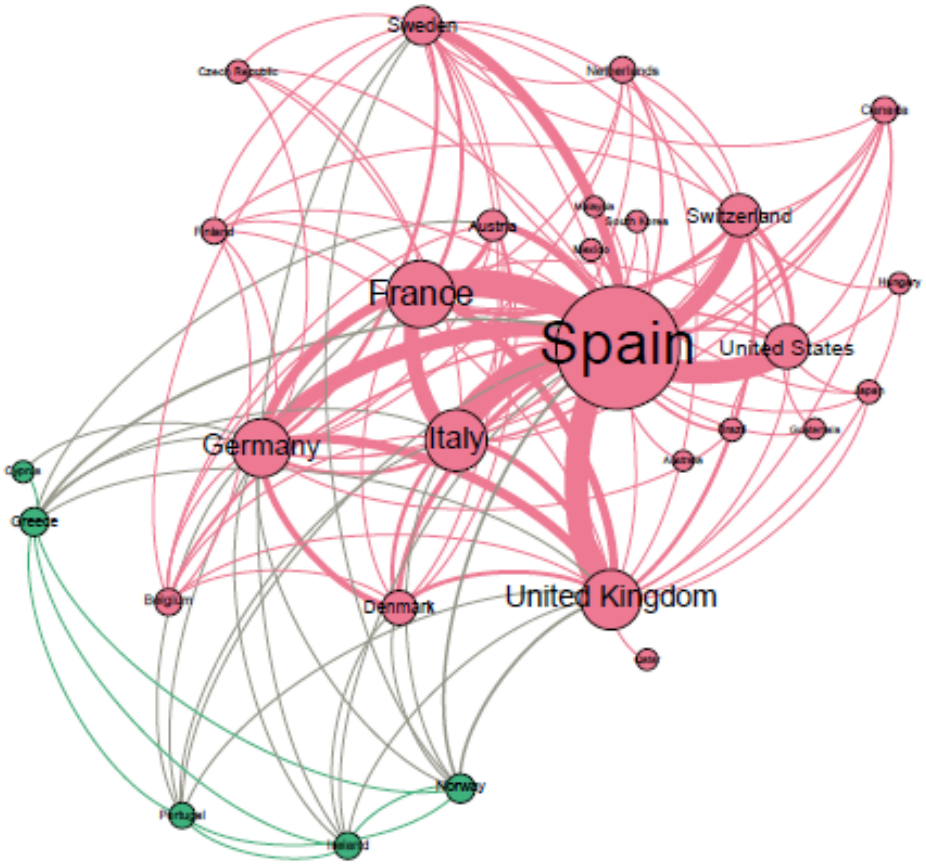


Figure 4: Scientific collaboration among countries.

In order to identify Iberdrola's scientific collaboration activities, an effective method is a network analysis. The network was plotted through Gephi software and the main affiliations that collaborate with Iberdrola were identified. The analysis has been carried out using matrix of co-occurrences and plotted in a network where the affiliations have at least one collaborative publication (see Figure 5).

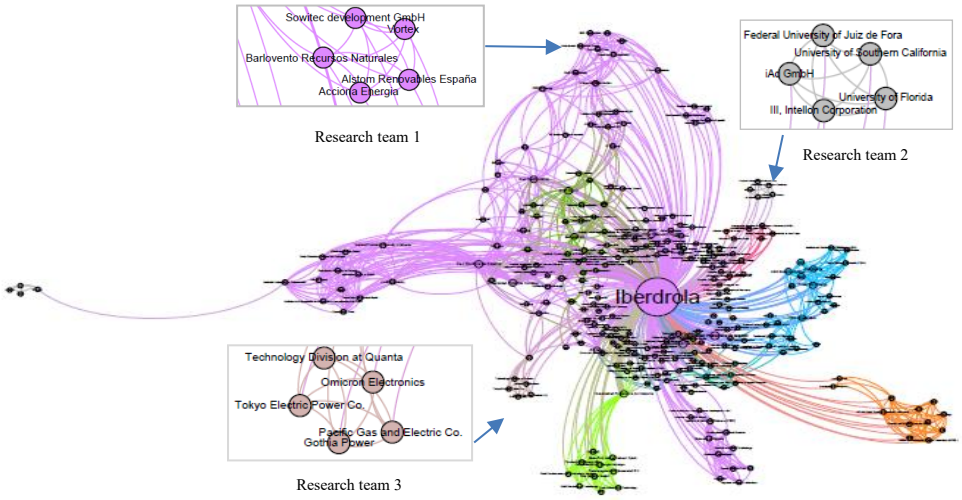


Figure 5. Scientific collaboration network of Iberdrola.

As shown in Figure 5, Iberdrola's scientific collaboration network is very extensive and productive. The network shows that Iberdrola collaborates with different research groups, formed by universities, prestigious research centers and major companies in the energy sector. The company is focusing its collaborative efforts on new renewable energy technologies, smart grids and smart mobility. Some of the small groups are analyzed below: research team 1 is formed by Alstom Renovables, Acciona, Vortex, etc. and researches issues related to renewable energy resources. Research group 2 is made up of the Universities of Florida, California, Juiz de Fora, Intellon and iAd GmbH, and their research topic is related to Electric power. The research stream of the third group selected (Omicron Electronics, Technology Division at Quanta, etc.) is operating practices with damaged equipment prevention.

3.2. Technological collaboration network

Technological collaboration is analyzed by using patent data. For this, the PatSeer database is used, being a global patent database. The search has been carried out on patent families. A patent family (PF) is a collection of patent applications covering the same or similar technical content. For the period from 1982 to 2019 a total of 131 PFs were detected.

In order to reflect inventive performance, the top assignee or applicants are from Spain and the USA, such as, Angel Iglesias S.A (Electronic and Communications: 14PFs), New York State Electric & Gas (NYSEG) (Energy supplier: 13PFs), Energetix (Energy supplier: 12 PFs) and Enertron S.A. (Renewable energies: 5PFs). Regarding the priority country, or the country where they were invented, Spain dominated with 84 PFs, the second country is the USA with 23 PFs, followed by Germany (8 PFs), Russia (6 PFs) and South Africa (4PFs).

As far as technological cooperation between applicants is concerned, an applicant collaboration network was generated and plotted with Gephi. During the process of cleaning up the patent database, six iberdrola entities have been identified: Iberdrola, Iberdrola Ingeniería y Consultoria, Iberdrola Ingeneria y Construccion, Iberdrola Generación Nuclear, Iberdrola Renovables Energia and Iberdrola Generación. As shown in Figure 6, the main entity in the collaboration network is Iberdrola, and the collaborative work takes place between energy sector companies and are clustered in small workgroups around the main node, which is Iberdrola. (Angel Iglesias S.A.- NYSEG – CSIC- Enertron; Energetix-Rochester gas electric- NASA; Arteche Hermanos- Escuela de Ingeniería Bilbao- Red Electrica Española), both national and international, which in turn have the most patents along with Iberdrola. In addition, it is worth mentioning the collaboration with scientific transfer entities such as the University of the Basque Country through the Faculty of Engineering in Bilbao, or the Higher Council for Scientific Research, or NASA. There are other subgroups within the main group that represent small clusters because they work with highly specific technology.

Regarding other entities of Iberdrola, Iberdrola Generación Nuclear and Iberdrola Ingeniería y Consultoría are beneficiaries of the patents without collaboration. However, Iberdrola Generación collaborates with Universidad Politecnica Madrid, and Iberdrola Ingeniería y Construcción collaborates with Iberdrola Renovables Energia and Scientific research centers, such as University of Salamanca and Coruña.

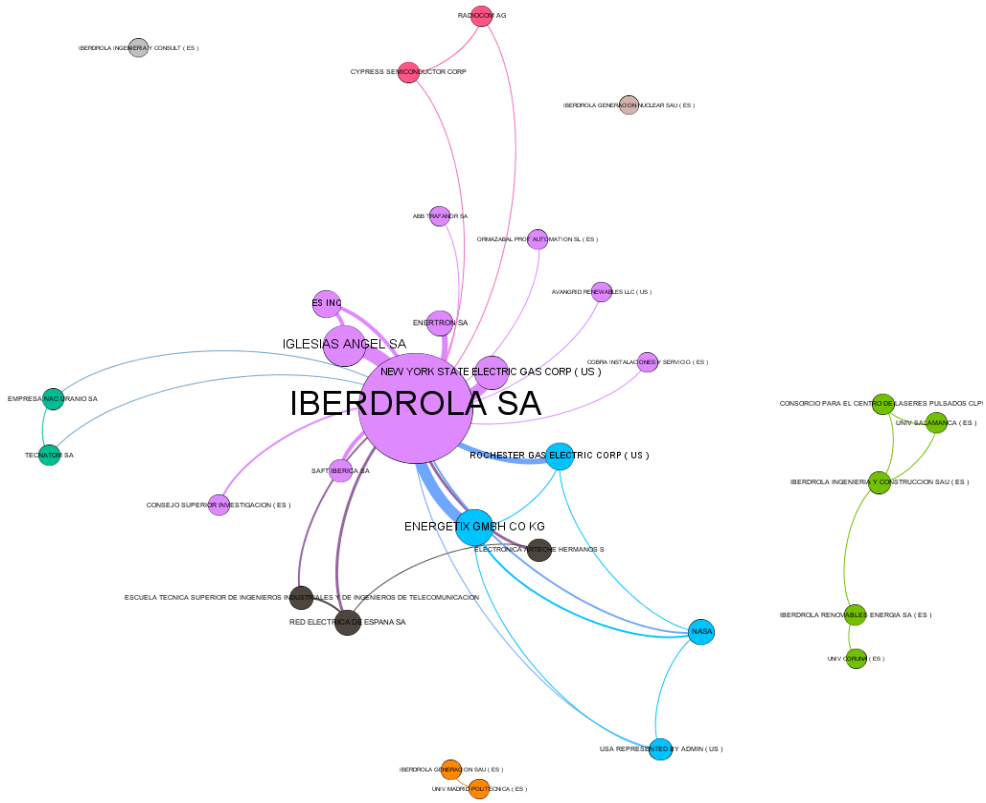


Figure 6. Technological collaboration network.

4. Conclusions

Open Innovation is an organizational model associated with Industry 4.0, which promotes knowledge transfer between different organizations. In the case of Iberdrola, which following its philosophy of Open Innovation, it shares its scientific knowledge with important entities from all over the world, clearly committed to internationalization in its collaborations. This active tendency to scientific development jointly with other entities allows Iberdrola to achieve its organizational objective linked to its philosophy of Open Innovation. In terms of technological collaboration, it can be said that Iberdrola shares much of the technological development carried out with other large companies in the energy sector. This is an important step, taking into account the industrial secrecy associated with patents, consequently, they have become "coopetitors", a term defined by Gnyawali and Park (Gnyawali & Park, 2011)

as the search for competition and collaboration between two or more companies at the same time. However, it is important to note that there is little collaboration with scientific entities, such as universities and technology centers, in the development of these innovations. It should also be noted that Iberdrola collaborates in both science and technology with entities such as the Red Electrica Española, the University of the Basque Country and NASA, among others.

References

Anderson, A. R., & Hardwick, J. (2017). Collaborating for innovation: the socialised management of knowledge. *International Entrepreneurship and Management Journal*, *13*(4), 1181–1197. https://doi.org/10.1007/s11365-017-0447-6

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. BT - International AAAI Conference on Weblogs and Social. *International AAAI Conference on Weblogs and Social Media*, 361–362.

Bildosola, I., Río-Bélver, R. M., Garechana, G., & Cilleruelo, E. (2017). TeknoRoadmap, an approach for depicting emerging technologies. *Technological Forecasting and Social Change*, *117*, 25–37. https://doi.org/10.1016/J.TECHFORE.2017.01.015

Bogers, M. (2011). The open innovation paradox: Knowledge sharing and protection in R&D collaborations. *European Journal of Innovation Management*, *14*(1), 93–117. https://doi.org/10.1108/14601061111104715

Chesbrough, H. W. (2003). *Open Innovation: The New Imperative for Creating and Profiting from Technology* (Harvard & B. S. Press, eds.). USA.

Elsevier. (2017). *Scopus: content coverage guide*. Retrieved from https://www.elsevier.com/__data/assets/pdf_file/0007/69451/0597-Scopus-Content-Coverage-Guide-US-LETTER-v4-HI-singles-no-ticks.pdf

Gnyawali, D. R., & Park, B. J. (2011). Co-opetition between giants: Collaboration with competitors for technological innovation. *Research Policy*, *40*(5), 650–663. https://doi.org/10.1016/j.respol.2011.01.009

Griliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature*, *28*, 1661–1707. https://doi.org/10.3386/w3301

Iberdrola. (2020). Open innovation and partnerships - Iberdrola. Retrieved from https://www.iberdrola.com/sustainability/innovation/open-innovation-partnerships

Leydesdorff, L. A. (2001). *The challenge of scientometrics : the development, measurement, and self-organization of scientific communications*. Universal Publishers.

Organisation for Economic Co-operation and Development. (2009). *OECD patent statistics manual.* Retrieved from http://www.oecd.org/sti/inno/oecdpatentstatisticsmanual.htm

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421–432. https://doi.org/10.1023/B:SCIE.0000034384.35498.7d

Sinha, M.; Pandurangi, A. (2016). *Guide to practical patent searching and how to use PatSeer for patent search and analysis* (2nd ed.). India: Gridlogics Technologies Pvt. Ltd.

Tejedor-Escobar, E., & Martínez-Cid, P. M. (2009). Red de Innovación de Iberdrola. *Dyna (Spain)*, *84*(6), 470–480.

Zemigala, M. (2019). Tendencies in research on sustainable development in management sciences. *Journal of Cleaner Production*, *218*, 796–809. https://doi.org/10.1016/j.jclepro.2019.02.009

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11611

Data granularity in mid-year life table construction

Jose M. Pavía[1], Natalia Salazar[2], Josep Lledó[3]
[1]Universitat de Valencia, Valencia, Spain, [2]Universidad Carlos III, Madrid, Spain, [3]Universidad de Alcalá, Alcalá de Henares, Spain.

Abstract

Life tables have a substantial influence on both public pension systems and life insurance policies. National statistical agencies construct life tables from death rate estimates (m_x)*, or death probabilities* (q_x)*, after applying various hypotheses to the aggregated figures of demographic events (deaths, migrations and births). The use of big data has become extensive across many disciplines, including population statistics. We take advantage of this fact to create new (more unrestricted) mortality estimators within the family of period-based estimators, in particular, when the exposed-to-risk population is computed through mid-year population estimates. We use actual data of the Spanish population to explore, by exploiting the detailed microdata of births, deaths and migrations (in total, more than 186 million demographic events), the effects that different assumptions have on calculating death probabilities. We also analyse their impact on a sample of insurance product. Our results reveal the need to include granular data, including the exact birthdate of each person, when computing period mid-year life tables.*

Keywords: *Mortality tables; mid-year estimators; death rates; big microdata; exposed-to-risk population.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Demographic processes affect how populations evolve over time. Understanding these processes is vital for a proper management of social systems, such as pensions and insurance schemes. Exploring mortality dynamics by exploiting all the data at hand is a key factor for improving death probability estimates. Information on variables such as death and birth rates or migratory flows is used by statistical bureaux to construct life tables, which are the cornerstone of the life insurance business.

Since the creation of the first life tables (Graunt, 1662), death probabilities (q_x) and death rates (m_x) have been estimated comparing deaths and exposed-to-risk. For death rates, the numerator refers to the number of deaths with age x in year t, D_x^t, while the denominator accounts for either the total number of 'person-years' at risk (e.g., Wilmoth et al. 2007; Arias, 2015; INE, 2016) or the average population at risk of dying, known as mid-year population estimates (e.g., ONS, 2012). In the era of the IT revolution and the boom of big data, the approach for computing the total number of 'person-years' at risk has been extensively studied in Lledó et al. (2019), while the approach for estimating mid-year figures remains under-researched.

Before the current overabundance of demographic microdata (Ruggles, 2014), period mid-year estimates required assumptions to be computed, such as a uniform distribution of births and deaths and, in some cases, the consideration of a closed population. Nowadays, these assumptions are becoming unnecessary. This paper (i) develops a period mid-year estimator unbounded by a set of hypotheses, (ii) studies (with a real dataset) the impact of using such an estimator on life table construction, and (iii) analyses its effect when calculating premiums for various insurance products.

The rest of the paper is organized as follows. Section 2 describes the methodology and presents the notation. In this section, five different mid-year estimators are discussed. Section 3 briefly introduces the dataset as well as the software used. Section 4 presents the main results, comparing the hypothesis-free estimator we introduce in section 2 with the classical mid-year estimator. In this section, we also analyse the impact of the hypothesis on an insurance product. Section 5 concludes and states future research questions.

2. Methodology

2.1. Preliminary Information

The methodology used counts the demographic variables (births, deaths, migrations, and emigrations) as expressed on the Lexis diagram (Lexi, 1880). Life tables are built based on period estimators, meaning they follow the population over a certain period of time, usually an even number of years. For instance, for construction of British life tables, demographic

events over 3 consecutive years are considered. In this paper, the count is based on the elements of the diagonal portion of the Lexis diagram corresponding to the areas ABFE and DEIH in Figure 1 for one-year tables and on the areas ABFE and DEIH for year 1, GHLK and JKON for year 2 and MNRQ and PQUT for three-year based tables. In both cases, depending on the type of estimator, different assumptions are made to determine which elements are relevant within the area of interest. The specific hypotheses and procedures are further explained in each section.

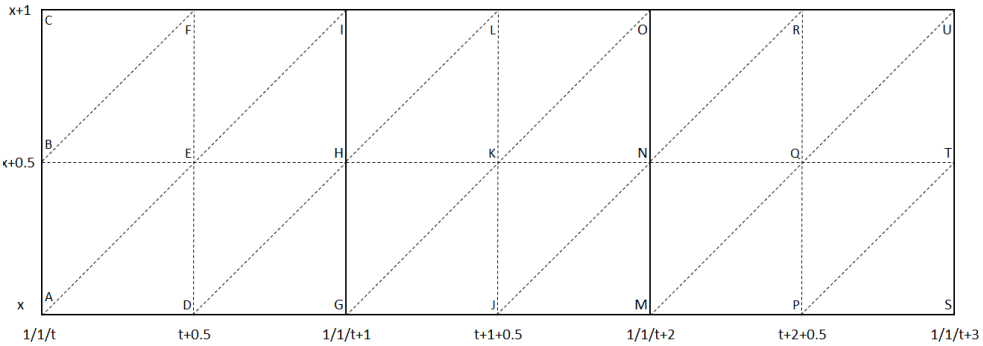


Figure 1. Lexis diagram.

Focusing on the one-year table, mortality rates link the amount of exposed population and deceased individuals through the formula: $\acute{m}_x = D_x^t / C_x^{t+0.5}$, where D_x^t refers to the number of deaths with completed age x in year t and $C_x^{t+0.5}$ to the population with completed age x in the middle of the year t.

We now consider the kind of (micro)data usually available in current statistical systems to develop different estimators by progressively relaxing the hypothesis for their construction until an (apparently) free-hypothesis estimator is reached.

2.2. Closed population and uniform distribution of deaths and birthdates (CP_UD_UB)

In the first scenario, we assume that deaths and birthdates are distributed in a uniform way across every year, implying that each of the triangles that form the Lexis diagram contains the same number of events of interest (1/8 of the total count within the year). Hence, for one period table:

$$\acute{m}_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\frac{1}{2}C_x^t - \frac{2}{8}D_x^t + \frac{1}{2}C_x^{t+1} + \frac{2}{8}D_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}} \qquad (1)$$

and generalizing for *i* periods:

$$\acute{m}_{x,i} = \sum_{j=1}^{i} \left(\frac{D_x^{t+j-1}}{\frac{1}{2} C_x^{t+j-1} + \frac{1}{2} C_x^{t+j}} \right) \tag{2}$$

To count the people who are alive with completed age x in the middle of year t, represented by segment DF, we start with those alive at the beginning of year t $(t+1)$, which corresponds to segment AB (HI). This segment equals half of segment AC (GI) due to the assumption of uniformity of birthdates, and corresponds to half of the people alive at age t with completed age $(x+1)$, denoted by C_x^t (C_x^{t+1}). The deaths occurring in triangles ABE (DEH) and BFE (EIH) are subtracted (added) to AB (HI) which results in segment EF (DE).

2.3. Open population and uniform distribution of deaths, migrants and births (OP_UD_UM_UB)

In this scenario, the distribution of deaths, birthdates and migratory flows are considered uniform across the year. In this case, the estimator results in the following extension of the basic mortality rate formula:

$$m_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\frac{1}{2} C_x^t - \frac{2}{8} D_x^t + \frac{2}{8} N_x^t + \frac{1}{2} C_x^{t+1} + \frac{2}{8} D_x^t - \frac{2}{8} N_x^t} = \frac{D_x^t}{\frac{1}{2} C_x^t + \frac{1}{2} C_x^{t+1}} \tag{3}$$

To count the number of people who are alive in the middle of the year, represented by segment DF, we start with those people alive at the beginning of year t $(t+1)$ corresponding to segment AB (GI) which, due to the assumption of the uniformity of births, is equal to half of the people with completed age $(x+1)$, C_x^t (C_x^{t+1}). By taking the difference between the total immigrations and emigrations occurring within the triangles ABE and BFE (DEH and EIH), the net migration flow, noted by N_x^t (N_x^{t+1}) is obtained which, as a result of the uniform distribution of both emigration and immigration, corresponds to 2/8ths of the net migration flow occurring across the whole year. The deaths along with the net migration occurring in the triangles ABE and BFE (DEH and EIH) are subtracted (added) to the count given by AB (HI), which results in segment EF (DE).

Since the net effects of migration for the first and second half of the year cancel out, the simplified version of the equation for this estimator is equal to the final formula of the previous one. The generalization for *i* periods stays the same as before: expression (2).

2.4. Closed population with no hypothesis about the distribution of deaths and uniform distribution of births (CP_NUD_UB)

As the hypothesis of uniformity for the distribution of deaths is relaxed, the same idea of section 2.2 is repeated, still taking segments AB and HI as half of the population with completed age x and $x+1$, respectively, but unlike estimator (1) the number of deaths occurring within areas ABFE and DEIH is not derived from the uniformity assumption. Instead, the exact amount of deceased people in each triangle ABE, BFE, DEH, and EIH is counted and then subtracted or added accordingly. The estimator results in the following extension of the basic mortality rate formula. For one period:

$$\check{m}_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\frac{1}{2}C_x^t - D_{x,L}^t + \frac{1}{2}C_x^{t+1} + D_{x,R}^t} \qquad (4)$$

generalizing for *i* periods:

$$\check{m}_{x,i} = \sum_{j=1}^{i}\left(\frac{D_x^{t+j-1}}{\frac{1}{2}C_x^{t+j-1} - D_{x,L}^{t+j-1} + \frac{1}{2}C_x^{t+j} + D_{x,R}^{t+j-1}}\right) \qquad (5)$$

where $D_{x,L}^t$ ($D_{x,R}^t$) is the number of deaths of people with age x during the first (second) half of year t of individuals born in the second (first) half of year $t-x-1$ ($t-x$). This corresponds to area ABFE (DEIH).

2.5. Open population with no hypothesis about the distribution of deaths and migration and uniform distribution of births (OP_NUD_NUM_UB)

When considering the possibility of migration flows, the same procedure of section 2.3 is repeated, still taking segments AB and HI as half of the population with completed age x and $x+1$, respectively. In this case, unlike estimator (4) the number of deaths along with the total net migration flow occurring within areas ABFE and DEIH is not derived using the uniformity assumption. Instead, the exact amount of deceased people and net migrants in each triangle ABE, BFE, DEH, and EIH are counted and then subtracted or added accordingly. Hence, the estimator is written as follows, for one period:

$$\ddot{m}_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\frac{1}{2}C_x^t - D_{x,L}^t - E_{x,L}^t + I_{x,L}^t + \frac{1}{2}C_x^{t+1} + D_{x,R}^t + E_{x,R}^t - I_{x,R}^t} \qquad (6)$$

generalizing for *i* periods:

$$\ddot{m}_{x,i} = \sum_{j=1}^{i}\left(\frac{D_x^{t+j-1}}{\frac{1}{2}C_x^{t+j-1} - D_{x,L}^{t+j-1} - E_{x,L}^{t+j-1} + I_{x,L}^{t+j-1} + \frac{1}{2}C_x^{t+j} + D_{x,R}^{t+j-1} + E_{x,R}^{t+j-1} - I_{x,R}^{t+j-1}}\right) \quad (7)$$

where $E_{x,L}^t$ and $E_{x,R}^t$ ($I_{x,L}^t$ and $I_{x,R}^t$) are the number of emigrations (immigrants) of people with age x during the first and second half of year t of individuals born during the second half of year $t-x-1$ and the first half of $t-x$, respectively. These are areas ABFE and DEIH.

2.6. Open population with no hypothesis about the distribution of deaths, migration, and births (OP_NUD_NUM_NUB)

Finally, when the hypothesis of the uniform distribution of birthdates is not assumed, the same procedure of section 2.5 is repeated, the only difference being that unlike estimator (6) the count for segments AB and HI is not estimated by means of the uniform hypothesis. Instead, the exact amount of people alive in AB and HI is counted using a summation function. This last estimator is synthesized as follows, for one period:

$$\ddot{m}_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\sum_{d=0}^{0.5} C_{x,d}^t - D_{x,L}^t - E_{x,L}^t + I_{x,L}^t + \sum_{d=0.5}^{1} C_{x,d}^t + D_{x,R}^t + E_{x,R}^t - I_{x,R}^t} \quad (8)$$

generalizing for *i* periods:

$$\ddot{\dot{m}}_{x,i} = \sum_{j=1}^{i}\left(\frac{D_x^{t+j-1}}{\sum_{d=0}^{0.5} C_{x,d}^{t+j-1} - D_{x,L}^{t+j-1} - E_{x,L}^{t+j-1} + I_{x,L}^{t+j-1} + \sum_{d=0.5}^{1} C_{x,d}^{t+j-1} + D_{x,R}^{t+j-1} + E_{x,R}^{t+j-1} - I_{x,R}^{t+j-1}}\right) \quad (9)$$

where $\sum_{d=0}^{0.5} C_{x,d}^{t+j-1}$ is the exact number of people with completed age x $(x+1)$ alive at the beginning of year t $(t+1)$, corresponding to the segment AB (HI).

3. Data and Software

The microdata used to make the computations was purchased from the Spanish National Institute of Statistics (INE). The database consists of the number of deaths, emigrations, immigrations and birthdates; the day, month and year of each event are specified as well as the gender of the individual. This information is available for the period 2005-2008. Overall, we processed, analysed and dealt with more than 180.15 million population inputs, 1.5 million death inputs, 0.7 million of emigrant inputs and 3.2 million of immigrant inputs,

covering both genders. In total, more than 186 million demographic events were studied individually. All the estimators were calculated using the statistical software R, version 3.6.1 (R Core Team, 2019).

4. Results

Throughout the previous section, we have gradually incorporated more detailed information into the mid-year mortality estimator until a hypothesis-free estimator is reached. In this section, using data from 2005 to 2007 to resemble the British mid-year estimator, we compare the most restricted three-periods version of the estimator (CP_UD_UB) with the three-periods hypothesis-free estimator (OP_NUD_NUM_NUB). The formulas for these two estimators correspond to equations (2) and (9), respectively, when $i = 3$.

As the accuracy of the estimator increases, so does the computation time. To check if it is worth exploiting the detailed microdata, we proceed to investigate issue (ii). To do so, we use the absolute relative discrepancy, $m_x - \ddot{m}_x \vee \quad /\ddot{m}_x$, as dissimilarity statistic and compare the life tables calculated using estimators (2) and (9) with $i = 3$. Figure 2 shows the differences.

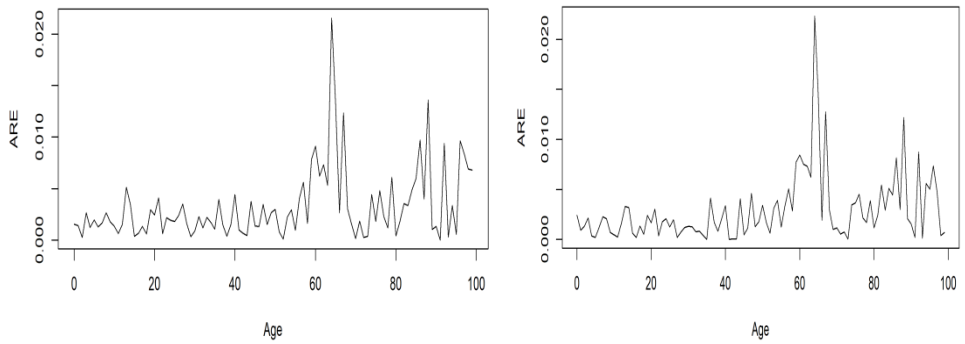


Figure 2. Absolute relative discrepancies between death rates computed using the CP_UD_UB and OP_NUD_NUM_NUB estimators for men (left panel) and women (right panel) for the period 2005-2007.

As Figure 2 shows, there are minimal differences until the age of 60 (around 1%). After this, two significant peaks emerge for the cohorts born just after the Spanish Civil War and just after World War II. Both occurrences cause a concentration of births during short time-spans that are not captured by the simplest form of the estimator (CP_UD_UB). The third peak, around the age of 90, is most likely a statistical effect due to the lack of population data for these ages, as fewer people tend to achieve this age.

To investigate issue (iii), we have analysed the impact that the different estimators have on an insurance product: in this case, a renewable year-term life insurance, with a sum insured of €100,000, for ages 50, 55, 60 and 65. To compute the premiums, we have assumed no

expenses and a discount rate of zero. The premiums are displayed in Table 1. Values for women are written between parentheses.

Table 1. Premiums payable for a renewable year-term life insurance of €100,000 for men (women). Years of study: 2005-2007.

Age	*CP_UD_UB*	*CP_NUD_UB*	*OP_NUD_NUM_UB*	*OP_NUD_NUM_NUB*
50	422.28 €	422.25 €	427.93 €	421.01 €
	(182.62 €)	(182.62 €)	(185.03 €)	(182 €)
55	637.28 €	637.21 €	645.66 €	636.66 €
	(246.14 €)	(246.14 €)	(249.07 €)	(245.84 €)
60	1,004.54 €	1,004.45 €	1,012.87 €	995.35 €
	(369.03 €)	(369 €)	(371.54 €)	(365.91 €)
65	1,430.16 €	1,429.33 €	1,447.76 €	1,410.84 €
	(562.58 €)	(562.44 €)	(570.03 €)	(554.41 €)

In general, the OP_NUD_NUM_NUB estimator premium is cheaper at ages approaching retirement for men and women. The biggest differences are 2.20% for men (2.28% for women) at age 64.

5. Conclusions and Future Research

In the actuarial and demographic fields, understanding mortality is always an important issue. The results obtained in this research have proven the impact of uniform hypothesis on both death probabilities and premium calculation. This has ramifications on best-estimate technical provisions under Solvency II and the new IFRS-17 regulatory frameworks. As demonstrated with the database analysed in this research, assuming a uniform distribution of birthdates is not appropriate. Hence, the estimator we propose in section 2.6 should be encouraged from a theoretical and practical perspective. Anyway, it would be interesting to compare for the whole range of ages all the statistical data contained in triangles BCF and DHG, for the annual estimator, and BCF and PTS, for the three-year estimator, as they have an impact on the numerator but not on the denominator of our estimators.

Acknowledgments

This research has been supported by the Spanish Ministry of Science, Innovation and Universities and the Spanish Agency of Research, co-funded with FEDER funds, project ECO2017-87245-R and Generalitat Valenciana (Consellería d'Innovació, Universitats,

Ciència i Societat Digital) project AICO/2019/053. The authors wish to thank Marie Hodkinson for revising the English text.

References

Arias, E. (2011). United States life tables. *National Vital Statistics Reports*, 64, 1-64.

Graunt, J. (1662). *Natural and Political Observations Made upon the Bills of Mortality*. London: Roycroft.

INE (2016). *Tablas de Mortalidad*, Madrid: Instituto Nacional de Estadística (Spain) [online]. Available at goo.gl/8Ywdc9.

Lexis, W. (1880). La representation graphique de la mortalité au moyen des points mortuaires. *Ánnales de Demographie Internationale*, 4, 297-324.

Lledó, J., Pavía, J.M., Morillas, F. (2019). Incorporating big microdata in life table construction: A hypothesis-free estimator. *Insurance: Mathematics and Economics*, 88, 138-150.

ONS (2012), "Guide to Calculating Interim Life Tables," Hampshire: Office for National Statistics (UK) [online]. Available at goo.gl/xFz7bV.

R Foundation for Statistical Computing (2019). Vienna. http://www.R-project.org/

Ruggles, S. (2014). Big microdata for population research. *Demography*, 69, 287-297.

Wilmoth, J. R., Andreev, K. F., Jdanov, D. A. and Glei, D. A. (2007). Methods protocol for the Human Mortality Database. *Human Mortality Database*. University of California Berkeley and Max Planck Institute for Demographic Research [online]. Available at http://www.mortality.org/.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11613

Extracting User Behavior at Electric Vehicle Charging Stations with Transformer Deep Learning Models

Daniel J. Marchetto[1], Sooji Ha[2, 3], Sameer Dharur[4], Omar Isaac Asensio[1,5*]
[1]School of Public Policy, Georgia Institute of Technology, United States, [2]School of Civil and Environmental Engineering, Georgia Institute of Technology, United States, [3]School of Computational Science and Engineering, Georgia Institute of Technology, United States, [4]School of Computer Science, Georgia Institute of Technology, United States, [5]Institute for Data Engineering and Science, Georgia Institute of Technology, United States.

Abstract

Mobile applications have become widely popular for their ability to access real-time information. In electric vehicle (EV) mobility, these applications are used by drivers to locate charging stations in public spaces, pay for charging transactions, and engage with other users. This activity generates a rich source of data about charging infrastructure and behavior. However, an increasing share of this data is stored as unstructured text—inhibiting our ability to interpret behavior in real-time. In this article, we implement recent transformer-based deep learning algorithms, BERT and XLnet, that have been tailored to automatically classify short user reviews about EV charging experiences. We achieve classification results with a mean accuracy of over 91% and a mean F1 score of over 0.81 allowing for more precise detection of topic categories, even in the presence of highly imbalanced data. Using these classification algorithms as a pre-processing step, we analyze a U.S. national dataset with econometric methods to discover the dominant topics of discourse in charging infrastructure. After adjusting for station characteristics and other factors, we find that the functionality of a charging station is the dominant topic among EV drivers and is more likely to be discussed at points-of-interest with negative user experiences.

Keywords: *Electric Vehicles; Mobility; Mobile Data; Natural Language Processing; Transformer Models.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

The transportation sector is undergoing rapid transformation such as vehicle electrification and increased usage of mobile apps. These two developments offer the possibility to do real-time monitoring of large-scale infrastructure with streaming data. Electric vehicles have also become a dominant strategy to reduce emissions that includes health co-benefits from displacing internal combustion engines (Carley et al., 2013; Sheldon et al., 2017). The growth of the EV market has brought an increase in complementary digital infrastructure—including charging stations and locator apps intended for use in public spaces. This digital infrastructure has created an ecosystem for users to engage with each other and share information about their EV experiences. The resulting user-generated data can be useful for policy analysis and real-time infrastructure management; however, large portions of this data is in the form of unstructured text, which require computational methods to extract insights (Asensio et al., 2020; Kühl et al., 2019).

In this article, we deploy recent advances in neural-net-based classification algorithms in order to learn the dominant topics of discourse within the EV community. This task of natural language processing (NLP) has been challenging because, although neural-net-algorithms have been shown to perform well in sentiment classification tasks, there are still computational issues related to underdetection particularly with highly imbalanced data (Asensio et al., 2020; Ha et al., 2020). Our approach here is to implement transformer based deep neural networks such as Bidirectional Encoder Representations from Text (BERT) and Transformer-XL (XLnet), which have both yielded promising results in a number of NLP tasks (LeCun & Hinton, 2015; Vaswani et al, 2017; Devlin et al., 2018; Yang et al., 2019). In order to understand user behavior in this domain, we begin with predefined behavioral topics as identified in Ha et al. (2020) and build multi-label classification models that assign one or more relevant topic labels to a given user text as demonstrated in Dharur et al. (2020). Using the output of these supervised classification architectures to find the dominant topics of discussion, we then implement econometric techniques to adjust the algorithmic predictions for observable station characteristics and other factors. We analyzed the multi-labeled topic classifications by points of interest (POI) to find evidence of charging station quality. We comment on implications for the use of transformer deep learning models in policy analysis and infrastructure management.

2. Data and Methodology

We have a nationally representative sample of unstructured consumer reviews at 12,720 U.S. charging station locations as provided by a popular EV charging station locator app. The text data consists of 127,257 reviews written in English from 29,532 registered and unregistered EV drivers during the period from 2011 to 2015, which are the early growth years of the EV

infrastructure. These represent charging station usage for the entire U.S. market during the period of study. In the sample, contemporaneous information from the mobile app was used to geocode the data with checks against Google Places API data. Example POI categories include Parking Garage/Lot, Government, Healthcare, Hotel/Lodging, Restaurant, School/University, Store/Retail, Workplace, etc. during the given year a review was made. For descriptive statistics of the review data by station, user and year, see Table 1.

Table 1. Review Count Descriptive Statistics.

Reviews per	Mean	SD	Min	Max
Station	9.26	21.22	1	578
User	5.17	15.96	1	728
Year (2011 – 2015)	28,034	22,024	1,331	50,217

2.1. Data Collection

As in many supervised machine learning tasks, we built a pre-labeled training set using the typology identified in Ha et al. (2020). There are 8 topics—Functionality, Range Anxiety, Availability, Cost, User Interaction, Location, Service Time, and Dealership—selected for human annotation of the review text. Reviews outside of these topics were labeled as Other. We recruited 5 annotators and provided a series of guidelines including a codebook with label definitions, examples from actual reviews for each topic, followed by a 1-hour guided training using a web application developed for annotation (Ha & Marchetto, 2020). After annotator training, inter-rater agreement (Fleiss' kappa) on a holdout sample for all topics ranged from 0.30 to 0.72. This calculation indicated substantial agreement for Service Time (0.72), Availability (0.61), Cost (0.65); moderate agreement for Range Anxiety (0.56), Functionality (0.52), Dealership (0.51), Location (0.45); and fair agreement for User Interaction (0.30). A total of 10,133 randomly selected, unique reviews were labeled by the 5 trained annotators. This selection is intended to be representative of the full dataset and includes reviews across all 8 main topics. Table 2 shows the counts and percentages of each labeled topic in the training data. The most frequently selected labels were Functionality, Location and Availability. We see the highest imbalance in Range Anxiety and Service Time, which were the least frequently selected labeled topics. Other represented only 1.1% of the training data.

Table 2. Counts and Percentages of Labeled Topics in Training Data.

Functionality	Location	Availability	Dealership	Cost	Service Time	Range Anxiety	Other
5,399	3,377	2,197	1,391	1,072	982	513	116
52.7%	33.0%	21.5%	13.6%	10.5%	9.6%	5.0%	1.1%

2.2. Classification through BERT and XLnet

Our classification task is to assign a predefined topic labels to a given text sequence. The advent of transformer-based deep neural network models has set new benchmarks on a wide variety of NLP tasks such as sentiment analysis, topic classification, question answering, machine translation, among others (Vaswani et al., 2017). A key reason for this algorithmic advancement was the use of the attention mechanism (Lin et al., 2017; Yang et al., 2016). This mechanism is a novel architecture that draws global dependencies between input and output and eliminates the need for other recurrent and convolution mechanisms. For more detailed expositions of BERT, XLnet and transformer models, see Vaswani et al. (2017), Devlin et al. (2018) and Yang et al. (2019). For the implementations described here, we followed the replication protocols outlined in Trivedi et al. (2019) and Dharur et al. (2020).

2.3. Fractional Response Models

Many observable station, location, and time factors can impact predictions of the topic classifications. Given possibilities for algorithmic bias from historical training data, we would also like to statistically adjust the algorithmic predictions to account for variations in use by POI, networks, and connector technologies available. Our unit of analysis is at the station level. The dependent variable $0 \leq y_{i,j,t} \leq 1$ is a standardized fractional response outcome (Equation 1) where a measure near 0 indicates a low incidence of a particular topic at that station location, while a score near 1 indicates a high incidence of a topic at that station location. This allows us to adjust our dependent variable for the usage frequency at a given station location. Given that Functionality is the dominant label in the training set, we focus our econometric analysis on that label. From Ha et al. (2020), Functionality refers to comments describing whether particular features or services are working properly at a charging station. Because we have a bounded outcome variable, we implement fractional response models (FRM) that use a quasi-maximum likelihood estimator (QMLE) to generate estimates of the likelihood of predicting a topic conditional on observable station characteristics. For additional details about FRM models, see Papke and Woolridge (1996) and Ramalho et al. (2011). The standardized topic score is,

$$y_{i,j,t} = Standardized\ Topic\ Score = \frac{Count\ of\ Topic\ Reviews_{i,j,t}}{Total\ Count\ of\ Reviews_{i,j,t}} \quad (1)$$

where *i* is a given review at *j* station location, in *t* year. Our main model specification is:

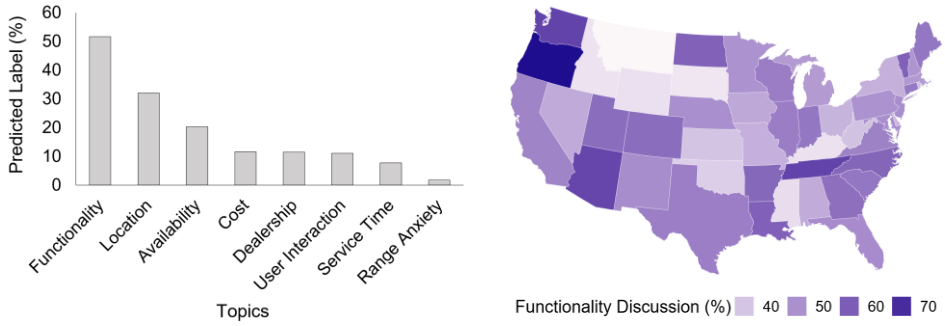
$$\begin{aligned} E(y_{i,j,t}|\boldsymbol{X_{i,j,t}}) &= G(\boldsymbol{X_{i,j,t}\beta}) \\ &= G(\beta_0 + POI\ Dummies_{i,j}\boldsymbol{\beta_1} + Station\ Characteristics_{i,j}\boldsymbol{\beta_2} \\ &+ Negativity\ Score_{i,j,t}\boldsymbol{\beta_3} + Interaction\ Effects_{i,j,t}\boldsymbol{\beta_4} \\ &+ Year\ Fixed\ Effects_i\boldsymbol{\beta_5}) \end{aligned} \quad (2)$$

where $\boldsymbol{X_{i,j,t}}$ is a vector of 1 by *k*, of explanatory variables, *E* is the expected value, and $G(\cdot)$ is a unit-bound, nonlinear transform function of a distribution defined as $\frac{e^{x_i\beta}}{1+e^{x_i\beta}}$, which is the logit function. The control variables, given in Equation (2) include POIs, Station Characteristics that include Number of Networks (e.g. 1, 2, 3+) and Number of Connector Types (e.g. 1, 2, 3) at a given station location. To mitigate the possibility for unobserved confounding variables, we also include a proprietary Station Quality Rating provided by the platform provider that ranges between 1 and 5, where 5 indicates a high-quality station location. The Negativity Score is a standardized measure of negative sentiment derived from Asensio et al. (2020). It is used to test the conjecture that negative experiences at different POIs differentially affect the likelihood that a review will be classified as Functionality through interaction effects. To calculate the partial effects and to assist with interpretation of the coefficients, the average effect on *y* of a unit change in $X_{i,j,t}$ estimated at the conditional mean is given by $\frac{\partial E(y_{i,j,t}|X_{i,j,t})}{\partial X_{i,j,t}} = \beta_i g(X_{i,j,t}\beta_i)$, where $g(X_{i,j,t}\beta_i)$ is $G(X_{i,j,t}\boldsymbol{\beta})[1 - G(X_{i,j,t}\boldsymbol{\beta})]$, estimated by the QMLE.

3. Results and Discussion

3.1. Classification

Across all topics of interest, we find that Functionality is the dominant topic among EV users (Figure 1). This is surprising because issues such as the cost of charging and range have typically received the most attention in public discourse on EV use. From Figure 1a, we see that Cost and Range Anxiety are not dominant topics of discussion. Surprisingly, in Figure 1b, we see that majority of the continental U.S. states discuss Functionality in over 50% of the reviews. Next, we report the classification results for accuracy (measured as partial accuracy) and F1 score (harmonious average of precision and recall). Table 3 contains the accuracy and F1 score of the overall multi-label topic classification. Accounting for uncertainty across 25 runs, we report a mean accuracy of 91.30% on BERT, 91.29% on XLnet, and a mean F1 score of 0.82 on both models. These results provide evidence that the use of these transformer-based models helped overcome the technical challenges of learning from imbalanced data (Table 2).



(a) (b)

Figure 1. (a) Frequency distribution of predicted topics across the entire dataset. (b) Percent of reviews discussing Functionality at the state-level.

Next, we evaluate the classification results using BERT and XLnet on each of the 8 topics of interest. In Figure 2, we report the accuracies achieved for each topic after 15 model replications. We compared this with a majority classifier, which predicts the most commonly occurring label by a simple majority. Improvements in accuracy over the majority classifier can be interpreted as a measure of the classifier's learning ability. From Figure 2, we see impressive accuracy improvement in the Functionality topic (23.3−23.7 percentage points), followed by Location (18.2−18.6), Availability (12.8−13.1), Cost (8.3−8.4), Dealership (6.2−7.6), Service Time (7.0−7.3) and user Interaction (3.9−4.1) topics. This result overcomes a common criticism of many machine learning algorithms with imbalanced data.

Table 3. Transformer model cross-validation results.

Architecture	Mean Accuracy % (s.d.)	Mean F1 Score (s.d)
BERT	91.30 (0.23)	0.82 (0.0071)
XLnet	91.29 (0.22)	0.82 (0.0046)

We find that the Range Anxiety topic gives the lowest accuracy improvement versus a majority classifier (0.4-0.6 percentage points). This could be due to this label being the least selected topic in our dataset, which suggests further scope for improvement to increase the size of the training data, or further tuning of the hyperparameters.

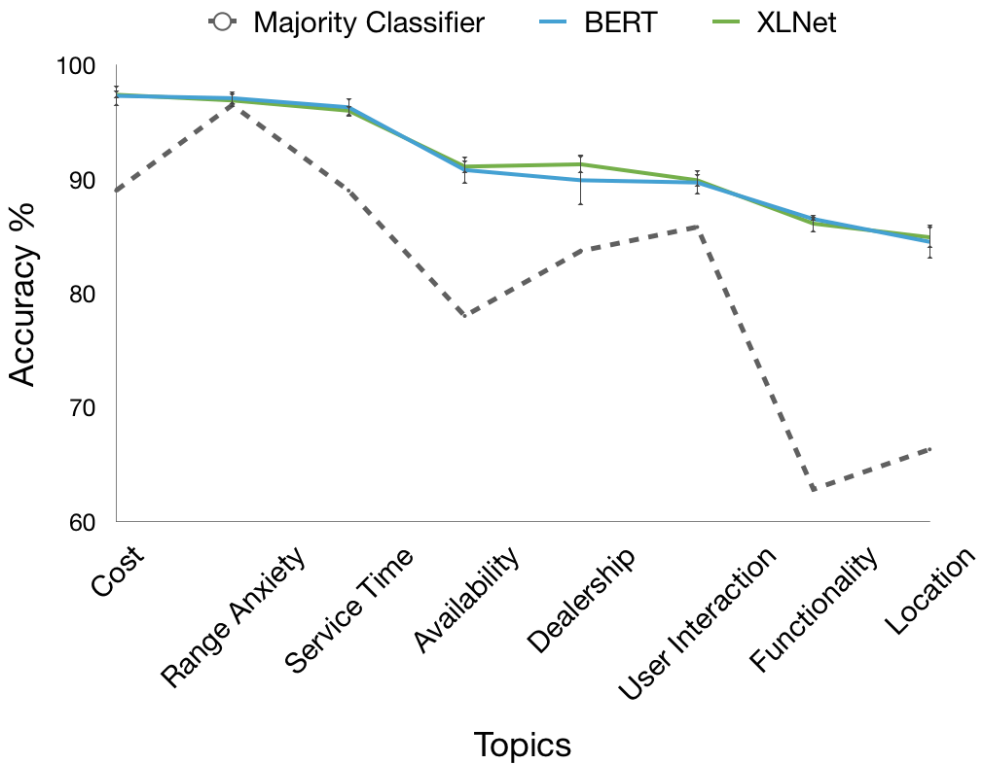


Figure 2. Topic-level accuracy comparisons with 95% C.I.

3.2. Fractional Response Models

In this section, we present the results for the fractional response models, which statistically adjust for station location and timing factors on the likelihood of selecting Functionality as the predicted label. Unlike many of other applications of machine learning, we argue that this statistical adjustment is needed to mitigate observational biases prevalent in training data. From Model 1 in Table 4, we find evidence of significant heterogeneity on the likelihood of discussing Functionality in public spaces. For example, compared with Street Parking and Parking Lots as the baseline counterfactual, consumers using stations located at POIs such as Shopping, Gas Stations, Supermarkets, Restaurants, and Hotels, are more likely to discuss the functionality of stations. These reviews typically discuss subtopics related to issues such as chargers, screens, connector types, connection, time, error messages, and customer service. However, POI locations such as Government, Healthcare, and Transit Stations were not statistically different from Street Parking and Parking Lots. In order to further understand if these discussions were related to negative user experiences, we evaluated the subpopulation of reviews by utilizing an algorithmically-generated sentiment score interacted with the POIs. In this analysis, we find that our most important and prominent POI, Shopping, was 7.4%

more likely to discuss Functionality issues as compared to our reference case (Model 3 in Table 3). Similarly, reviews at Hotels, although not significantly different from Street Parking and Parking Lots, were 6.4% more likely to discuss Functionality in the presence of negative sentiment. These results from consumer data could suggest that charging station operators at many public spaces or POIs may not have sufficient incentives to ensure the proper maintenance and upkeep of publicly accessible EV infrastructure. Future work could further evaluate mechanisms of dissatisfaction.

4. Conclusion

In this article, we have demonstrated the use of neural-net-based classification algorithms to automatically discover topics of EV discourse among members of the EV community. We provide evidence that transformer-based models overcome prior challenges of training models with highly imbalanced data. In the context of EV reviews, we then use these classification results to identify major issues that users experience in public charging infrastructure. We find that Functionality is the dominant topic of discussion with significant heterogeneity by POIs. This is counter to the public discourse that focuses on cost and range anxiety as dominant themes. This research also provides a proof-of-concept for large-scale practical implementation that can enable real-time processing of mobility behavior patterns.

Acknowledgments

We gratefully acknowledge funding by the National Science Foundation (NSF Award No. 1931980) and Microsoft Azure. We also thank IDEaS, and the PACE high performance computing team at Georgia Tech.

Table 4. Partial Effect Results from FRMs of Standardized Functionality Score.

	Model 1		Model 2		Model 3	
	Coef	**SE**	**Coef**	**SE**	**Coef**	**SE**
POIs						
Shopping	0.076**	0.022	0.048**	0.018	0.013	0.023
Car Dealership	0.060**	0.019	0.021	0.016	0.021	0.016
Workplace	0.081**	0.027	0.038	0.020	0.03	0.020
Gas Station	0.165**	0.021	0.167**	0.017	0.154**	0.019
Government	0.028	0.025	0.036	0.023	0.036	0.023
Supermarket	0.196**	0.028	0.149**	0.023	0.115**	0.037
Hotel	-0.006	0.022	0.000	0.019	-0.023	0.019
Restaurant	0.077**	0.028	0.084**	0.022	0.056*	0.027
Education	0.066*	0.027	0.049*	0.020	0.047*	0.022
Transit Station	0.035	0.034	0.038	0.027	0.038	0.027
Healthcare	-0.073	0.043	-0.001	0.028	-0.001	0.028
Entertainment	-0.031	0.034	-0.016	0.034	-0.015	0.034
Airport	0.040**	0.033	-0.114*	0.040	-0.129	0.077
Library	-0.007*	0.034	0.076*	0.034	0.077*	0.034
Residential	-0.023	0.034	0.098	0.067	0.099	0.067
Quality Rating			-0.018**	0.003	-0.018**	0.003
Negativity Score			0.130**	0.009	0.107**	0.011
Interactions of Negativity with:						
Shopping					0.074*	0.038
Gas Station					0.029	0.027
Supermarket					0.068	0.057
Hotel					0.064*	0.040
Restaurant					0.033	0.119
Education					0.058	0.028
Airport					0.006	0.033
Station Characteristics	Yes		Yes		Yes	
Year FE	Yes		Yes		Yes	
Clustered SE	Yes		Yes		Yes	
No. Observations	127,257		127,257		127,257	

*p < 0.05, **p < 0.01, ***p < 0.001

References

Asensio, O. I., Alvarez, K., Dror, A., Wenzel, E., Hollauer, C., & Ha, S. (2020). Real-time data from mobile platforms to evaluate sustainable transportation infrastructure. *Nature Sustainability.* DOI 10.1038/s41893-020-0533-6

Carley, S., Krause, R. M., Lane, B. W., & Graham, J. D. (2013). Intent to purchase a plug-in electric vehicle: A survey of early impressions in large US cities. *Transportation Research Part D: Transport and Environment*, 18, 39-45.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (pp. 4171-4186).

Dharur, S., Ha, S., Marchetto, D. J., & Asensio, O. I. (2020) Topic classification of electric vehicle consumer experiences with transformer-based deep learning. Working paper.

Ha, S., Marchetto, D. J., Burke, M. E., & Asensio, O. I. (2020) Detecting behavioral failures in emerging electric vehicle infrastructure using supervised text classification algorithms. In *Proceedings of the Transportation Research Board Annual Meeting.* 20-03461.

Ha, S. & Marchetto, D. J. (2020). Codebook, Available online at https://github.com/asensio-lab/transformer-EV-topic-classification/tree/master/training-manual.

Kühl, N., Goutier, M., Ensslen, A., & Jochem, P. (2019). Literature vs. Twitter: Empirical insights on customer needs in e-mobility. *Journal of Cleaner Production*, 213, 508-520.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, 11(6), 619-632.

Ramalho, E. A., Ramalho, J. J., & Murteira, J. M. (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, 25(1), 19-68.

Sheldon, T. L., DeShazo, J. R., & Carson, R. T. (2017). Electric and plug-in hybrid vehicle demand: lessons for an emerging market. *Economic Inquiry*, 55(2), 695-713.

Trivedi, K. (2019). Fast-Bert, Accessed online on 02/21/2020 at https://github.com/kaushaltrivedi/fast-bert.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Association for Computational Linguistics: Human Language Technologies conference* (pp. 1480-1489).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *In Advances in Neural Information Processing Systems* (pp. 5754-5764).

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11616

Comparative multivariate forecast performance for the G7 Stock Markets: VECM Models vs deep learning LSTM neural networks

Diana Mendes[1], Nuno Ferreira[1], Vivaldo Mendes[2]
[1]Department of Quantitative Methods for Management and Economics, ISCTE Business School, ISCTE-IUL, Portugal, [2]Department of Economics, ISCTE-IUL, Portugal.

Abstract

The prediction of stock prices dynamics is a challenging task since these kind of financial datasets are characterized by irregular fluctuations, nonlinear patterns and high uncertainty dynamic changes.

The deep neural network models, and in particular the LSTM algorithm, have been increasingly used by researchers for analysis, trading and prediction of stock market time series, appointing an important role in today's economy.

The main purpose of this paper focus on the analysis and forecast of the Standard & Poor's index by employing multivariate modelling on several correlated stock market indexes and interest rates with the support of VECM trends corrected by a LSTM recurrent neural network.

Keywords: *Stock Markets; multivariate forecasting; VECM; LSTM.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

Over the years, the usage of standard econometric practices has proven its usefulness. Several models, such as ARIMA (AutoRegressive Integrated Moving Average), VAR (Vector AutoRegression) and VECM (Vector Error Correcting Model), have been employed to analyze financial and macroeconomic variables in order to forecast or to extract dependencies and different causality relations between them.

One of the most popular econometric models, according to (Mills & Markellos, 2008) are error-correcting mechanisms. These models stem from the idea that common time series can have a long-term dependency on a stochastic trend. One major advance in this area came from Granger's representation theorem (Engle & Granger, 1987), which shows precisely that a cointegration relation can be represented by the error correction model (ECM).

In addition, given the recent popularity of machine learning techniques in econometrics, LSTM (Long Short-Term Memory) models have been engaged to financial indexes in order to try to predict market fluctuations. Examples of this are (Nelson, Pereira, & Oliveira, 2017) which use LSTM to predict BOVESPA (São Paulo Stock Exchange) within a 15 minute time window. They report accuracies of 53–55%. Another approach was that of (Fischer & Krauss, 2017) which studied daily Standard & Poor's data from 1992 to 2015.

Recently there has been an insurgence of hybrid models which combine both traditional econometric and time series methods with machine learning algorithms. Authors such as (Terui & van Dijk, 2002), (Zhang, 2003) and (Zhang & Qi, 2005) explore the idea that a single model can't fully recognize the true data generating process or identify all the characteristics of a time series.

(Kim & Won, 2018) propose a new hybrid model to forecast KOSPI 200 stock price volatility by combining LSTM recurrent networks with various generalized autoregressive conditional heteroscedasticity (GARCH)-type models. A different approach was that of (Wan, Guo, Yin, Liang, & Lin, 2020) which proposed a brand new LSTM hybrid model called CTS-LSTM. The purpose os this model is to forecast correlated time series by capturing any complex non-linear patterns existing between the variables.

In this paper we propose a multivariate approach which firstly analysis the overall trend of the considered data by using a VECM and lately corrects the non-captured patterns by using a LSTM (Long Short-Term Model) neural network. The VECM model is applied to capture the linearity in the original data and the resulting residuals are used as input for the LSTM algorithm with the purpose to extract nonlinear behaviour and to complete prediction.

2. Modeling the dataset by using VECM and LSTM algorithms

The main goal of this analysis is to infer the accuracy of long term predictions using both VECM and LSTM models but also to get benefit from each techniques by displaying what they do the best, namely, trend estimations (VECM) and pattern recognition (LSTM).

Our model philosophy is that of filtering the data, followed by the interpretation of the underlying patterns. We achieve this by using a VECM model to assert and predict the overall trend of the data and then apply an LSTM network to the VECM residuals.

2.1. Data handling and VECM fitting

The dataset we are going to use includes: Dow Jones, NASDAQ and Standard & Poor's 500 market indexes and the 3 month Treasury bill rate, sampled weekly, as can be observed in Figure 1. The data were cut short in time domain so that every series spaned over the same time period. All market indexes are the closing price on every Friday of the month and luckily the treasury bill rate was also available on the same day so no resampling or data shifting was required.

After collecting the data from the Reuter-Thomson Datastream platform (Reuters-Thomson, 2020) we test all time series for stationarity by using the Augmented Dickey–Fuller unit root test (ADF) (Dickey & Fuller, 1979; Dickey & Fuller, 1981). This test indicates that the four time series are non-stationary with p-values well above 5% significance level (0.9945, 0.9989, 0.9959, 0.6274). The log-returns are stationary (based on ADF test), which is conducent to a VECM analysis, on the assumption that cointegration exists. From this point on, all numerical results presented in the text in the form of 4-tuples are related to each of the variables in the order presented in Figure 1.

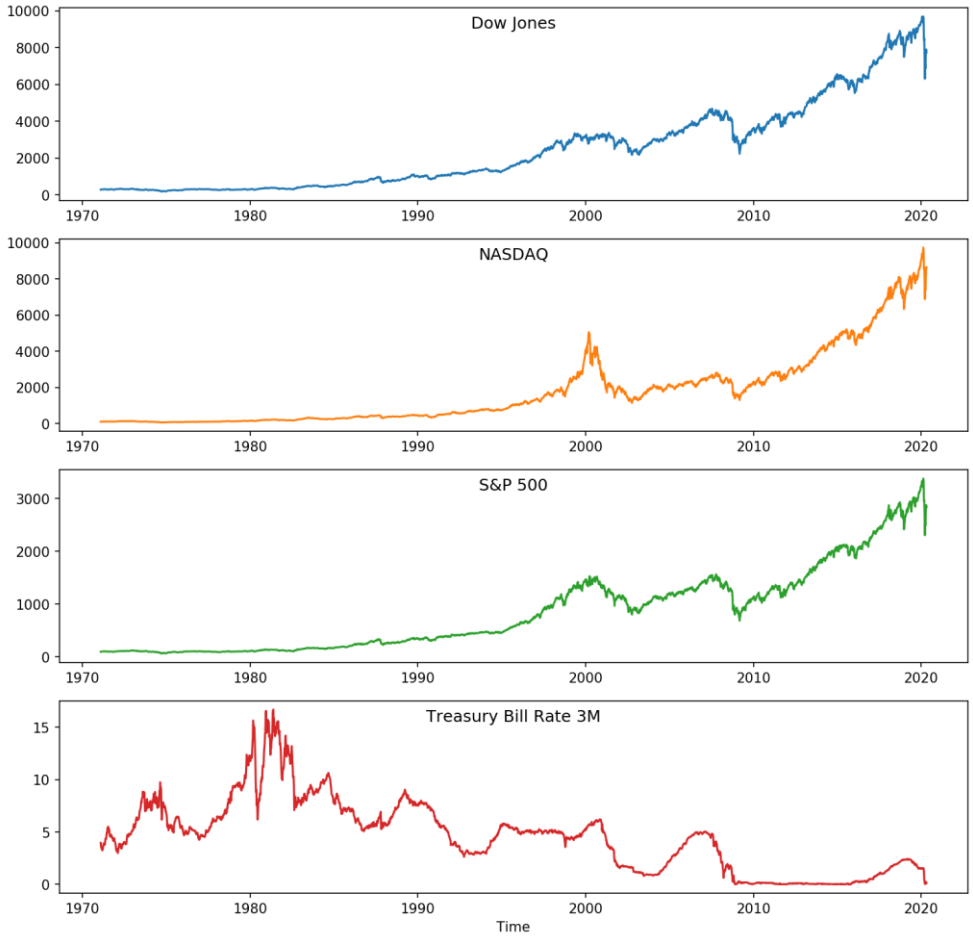


Figure 1. Macroeconomic variables and stock indexes in study.

We establish two major datasets: the training set for the VECM estimation and a test set which we use to benchmark VECM's prediction capabilities as well as to supply the input for the LSTM algorithm to effectively "correct". The training dataset runs from 5 of February 1971 until 12 of May 2013 while test set goes from 19 of May 2013 until 17 of April 2020.

In the VECM model fitting procedure we minimize Akaike Information Criteria (AIC) in order to determine the optimal lag, which in this case is of 2 time steps (weeks in our time binning). The cointegration was detected by adopting the Johansen methodology (Johansen, 1988; Johansen, 1992) where both trace and maximum eigenvalue tests conclude that exist three cointegration vectors for the optimal lag order of 2.

After the model was established we use the VECM(2) with 3 cointegration vectors to estimate the parameters and to extract the residuals (for the training set) and to forecast (for the test set). Note that these residuals were then scaled in order to comply with the LSTM network which will use them as input. These results can be found in Figure 2.

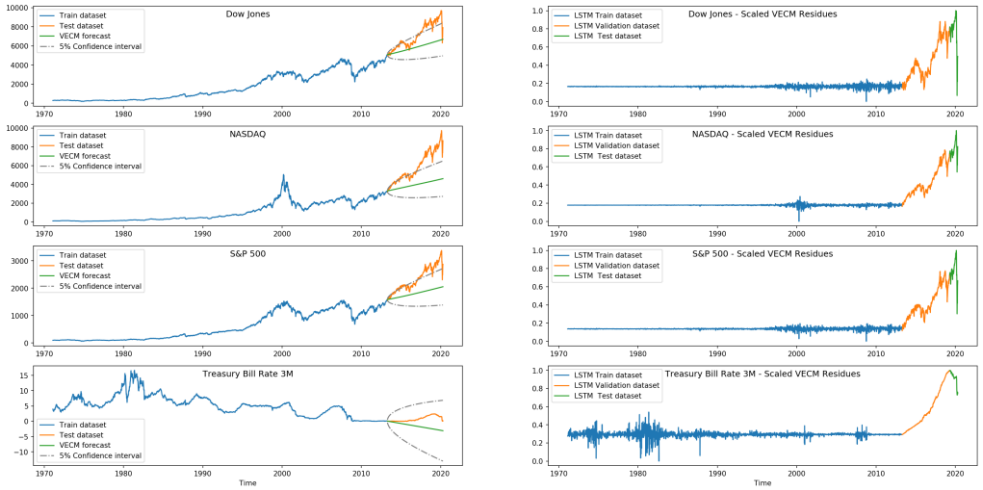


Figure 2. VECM model applied to data (left) and the corresponding residuals (right).

We estimated an average percentual deviation of the VECM model from data in the training sample given by (1.489%, 0.901%, 2.234%, 1.196%). The prediction results based on VECM, ilustrated in Figure 2, strongly deviate from the data as time passes but this was to be expected given the extremely long prediction range (about 7 years), resuming to only predict the global trend.

2.2. LSTM network

After the residuals were extracted and scaled to a 0 to 1 grading, we introduce them as input in our LSTM network.

We split the "forecast" dataset of the VECM into two subsets in order to create a "validation" and a new "test" datasets for the LSTM's training and benchmarking. The new time intervals splits are given by: 1971-02-19 to 2013-04-12 for training, 2013-04-19 to 2019-04-12 for validation and 2019-04-19 to 2020-04-17 for testing.

This network's configuration, presented in Figure 3 (namely the number of time steps, batch size, training epochs and hidden layer configuration) came from a long run of trial and error.

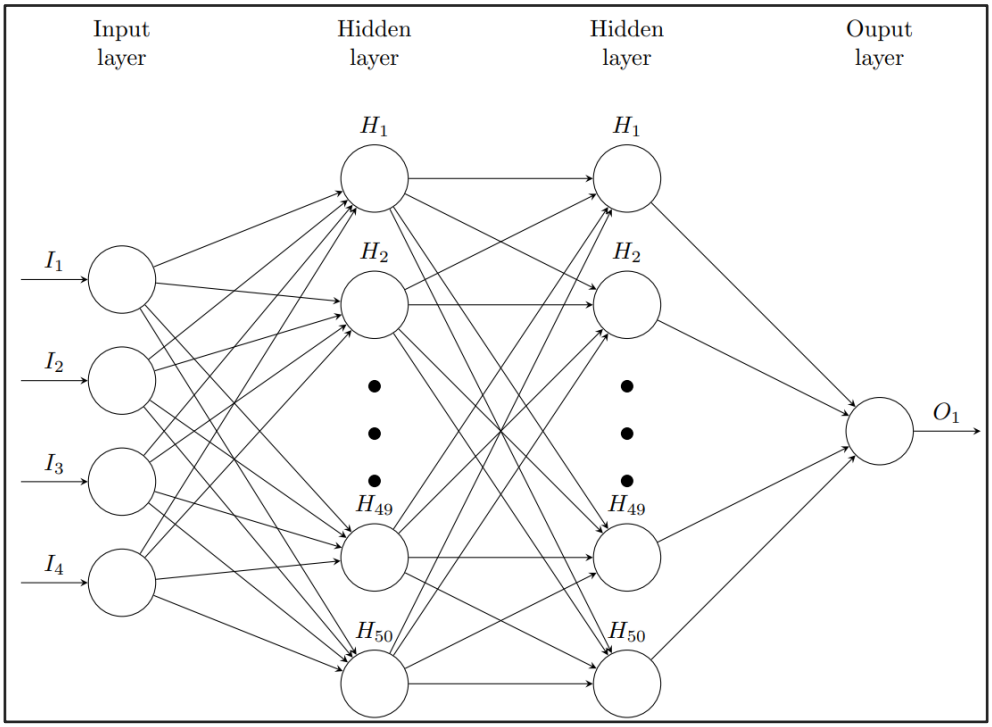


Figure 3. Diagram of the LSTM Recurrent Neural Network deployed.

Table 1 synthesize the different parameters that define our LSTM arquitecture. We use Keras and Tensorflow from Python. In order to normalize the data we use the MinMaxscaler, which scales and translates each feature individually such that will belong to the given range on the training set, e.g. between zero and one. To avoid using Sigmoid functions, ReLU (Rectified Linear Unit) activation functions became a popular choice in deep learning and even nowadays provides outstanding results. An optimizer is one of the two arguments required for compiling a Keras model. We used Adam, an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. Finally, we present the values exploit for the number of time steps, hidden layers, learning rate, batch size and training epochs.

Table 1. LSTM architecture.

Data normalization	MinMaxScaler
Activation function	ReLU
Optimizers	Adam
Loss Function	Mean Squared Error
Input dimension	4 (timestep) x 4
Output dimension	1 (forecast)
Hidden layers	[50, 50]
Learning rate	1.E-3
Batch Size	32
Training epochs	30

2.3. Results

After training, the network was used to predict the residuals of the time series in order to correct the VECM forecasts. This was merely done by adding the LSTM prediction to the VECM prediction after the scaling was inverted.

The results are presented in Figure 4. The average percentual deviation of this approach from the data is 28.192%.

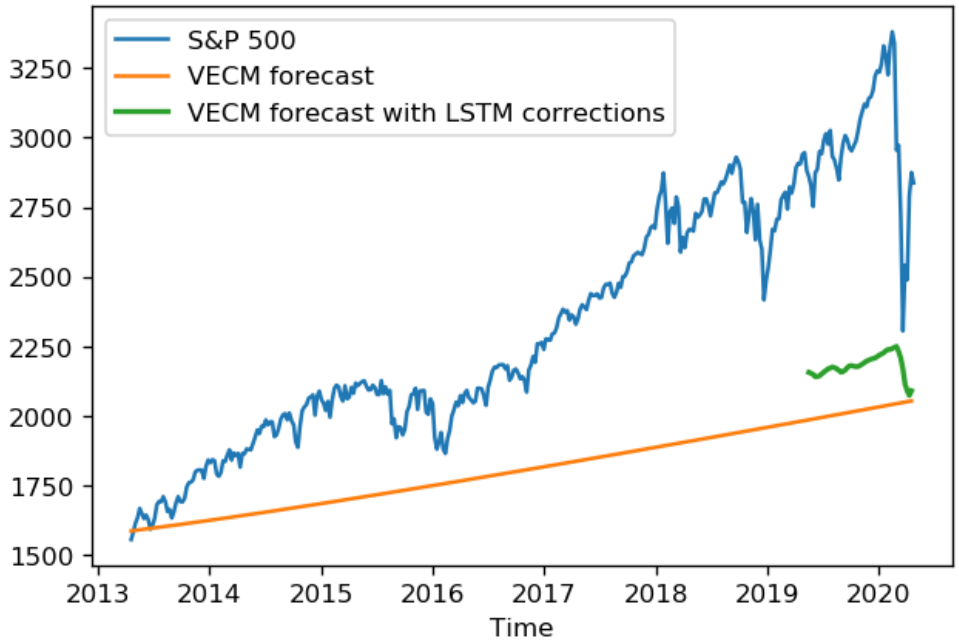


Figure 4. VECM prediction and the VECM with LSTM correction.

The LSTM's additive correction introduced some temporal variation but it was not able to correct for such a large shift from the forecast to the data. This shortage can be attributed to the discrepant change in the behaviour of the residual dataseries presented in Figure 2.

3. Conclusions and Prospects

The approach of making an unbiased long term "trend" prediction of the different time series and then attempting to correct them using an LSTM network proved to be difficult.

The residual series sharply change and the LSTM was not able to take that into account during its training since only a simple residual series was provided. This conduces to the introduction of two new objectives in future analysis, namely, an adjustment in the forecasting range and a change in the training philosophy.

This model yielded a MAPE of 28.192% when prediciting the Standard & Poor's stock market index. This might be attributed to the usage of a relatively small dataset.

We plan to expand this analysis to daily data but that will constrain our choices due to data availability of macroeconomic variables on these high frequency samplings.

References

Dickey, D. A., & Fuller, W. A. (1979, June). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association, 74*(366). doi:10.2307/2286348

Dickey, D. A., & Fuller, W. A. (1981, Jul). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica, 49*(4), pp. 1057-1072. doi:10.2307/1912517

Engle, R. F., & Granger, C. W. (1987, Mar). Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica, 55*(2), pp. 251-276. doi:10.2307/1913236

Fischer, T., & Krauss, C. (2017). Deep learning with long short-term memory networks. *FAU Discussion Papers in Economics*. doi:10.1016/j.ejor.2017.11.054

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control, 12*(2-3), pp. 231-254. doi:10.1016/0165-1889(88)90041-3

Johansen, S. (1992, June). Cointegration in partial systems and the efficiency of single-equation analysis. *Journal of Econometrics, 52*(3), pp. 389-402. doi:10.1016/0304-4076(92)90019-N

Kim, H., & Won, C. H. (2018, Aug). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications, 103*, pp. 25-37. doi:10.1016/j.eswa.2018.03.002

Mills, T. C., & Markellos, R. N. (2008). *The Econometric Modelling of Financial Time Series.* Cambridge University Press. doi:10.1017/CBO9780511817380

Nelson, D. M., Pereira, A. C., & Oliveira, R. A. (2017, May). Stock market's price movement prediction with LSTM neural networks. *International Joint Conference on Neural Networks (IJCNN)* (pp. 1419-1426). IEEE. doi:10.1109/IJCNN.2017.7966019

Reuters-Thomson. (2020). Datastream. Retrieved from https://www.refinitiv.com/en/products/datastream-macroeconomic-analysis/

Terui, N., & van Dijk, H. K. (2002). Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*(18), pp. 421-438. doi:10.1016/S0169-2070(01)00120-0

Wan, H., Guo, S., Yin, K., Liang, X., & Lin, Y. (2020, March 5). CTS-LSTM: LSTM-based neural networks for correlatedtime series prediction. *Knowledge-Based Systems, 191*. doi:10.1016/j.knosys.2019.105239

Zhang, G. P. (2003, Jan). Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing, 50*(17), pp. 159-175. doi:10.1016/S0925-2312(01)00702-0

Zhang, G. P., & Qi, M. (2005, Feb). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research, 160*(2), pp. 501-514. doi:10.1109/TNN.2007.912308

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11619

Investigating inefficiencies of bookmaker odds in football using machine learning

Benedikt Mangold[1], Johannes Stübinger[2]
[1]GfK, Germany, [2]Department of Statistics and Econometrics, University of Erlangen-Nuremberg, Germany.

Abstract

The efficient-market hypothesis states that it is impossible to beat the market, as the price reflects all available information. Applied to bookmaker odds for football games, there should not be a systematic way of winning money on the long run. However, we show that by using simple machine learning models we can systematically outperform the markets belief manifested through the bookmakers odds. The effect of this inefficiency is diminishing over time, which indicates that the knowledge that has been derived from and the pure amount of the data is also reflected in the odds in recent times.

We give some insights how this effect differs across major football leagues in Europe, which algorithms are performing best and statistics on the ROI using machine learning in football betting. Additionally, we share how the simulation study has been designed in more detail.

Keywords: *Machine Learning; Football.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

What if you could beat the market? Many tried, many failed. This paper analyses the efficiency of the market manifested in bookmaker odds in the domain of football. For that, we use publicly available information about football clubs and their active players to train a machine learning model that predicts the outcome of a match. The simulation covers the five major European football leagues with corresponding second leagues and data from twelve seasons. We use the predictions to (virtually) place a bet on historic odds to analyze if we could systematically outperform the believes of the market, measured by the return of the placed bet. In the following, we revisit the results of the paper by Stübinger et al. (2020) with respect to the success of the betting strategy, the implications for the efficiency of the offered odds and some detailed analyses from the perspective of the bookmakers during the covered period.

The first section revisits the simulation design and important results of Stübinger et al. (2020), the second section discusses the findings with respect to the well-known market efficiency hypothesis. The third section concludes this manuscript.

2. Modelling a football match with machine learning

Our simulation study is mainly based on two data sources we crawl from the internet. First, we consider all football matches from Primera Division, Segunda Division (Spain), Premier League, Football League Championship (England), Bundesliga, Bundesliga 2 (Germany), Serie A, Serie B (Italy), Ligue 1, Ligue 2 (France) from season 2006/2007 to 2017/2018. This record of 47,856 football matches provides a true hardness test for any back-testing study, as investor interest and analyst scrutiny are particularly high for these football nations. Second, we take into account 40 features for each player who was active in the respective matches. To be more specific, we consider skills from the areas “General”, “Ball Skills”, “Passing”, “Shooting”, “Defense”, “Physical”, “Mental”, and “Goalkeeper”.

In the spirit of Stübinger and Knoll (2018), the data set is sliced into 12 overlapping study periods, each shifted by one season. Each study period consists of a formation period and an out-of-sample trading period. The formation period identifies complex relations between the features of the players and the corresponding match result by fitting different machine learning models Random Forest (RAF), Boosting (BOO), Support Vector Machine (SVM), Linear Regression (LIR). Furthermore, we introduce a weighted ensemble method (ALL) by integrating the information of the four baseline approaches. The trading period (1) predicts football matches with the help of the mentioned data-driven methods and (2) exploits the obtained information in order to construct a statistical arbitrage strategy. If our assumptions hold, the trading algorithm would be able to find market anomalies and to generate positive profits.

Table 1 provides an overview of the results achieved by the different machine learning models. The ensemble strategy ALL results in the highest accuracy of 81.77% and an average payoff with a value of 1.0158. Note that with an average payoff that is greater than 1 one can beat the bookmaker over the long term as for each monetary unit spend, on average the return is strictly positive. Overall, the higher the complexity of a strategy, the higher the quality of our predictions which is reflected in higher average payoffs. It should be mentioned that we benchmark these strategies against baseline betting algorithms, e.g., randomly betting or always betting on the event with the lowest odd (most probable outcome) or placing bets on the home team. All these benchmarks perform significantly worse than strategies based on ML. We carefully conclude that (a) player characteristics contain information about the outcome of football matches and (b) our ML methods can capture and exploit these signals from the data.

Table 1. Statistical performance indicators for the betting strategies for the football seasons 2006/2007 to 2017/2018.

Key figure	RAF	BOO	SVM	LIR	ALL
Accuracy	0.8126	0.7912	0.6971	0.7292	0.8177
Average Payoff	1.0043	1.0072	0.9757	0.9933	1.0158

Source: Stübinger et al. (2020).

Figure 2 analyzes the performance of the strategies RAF, BOO, SVM, LIR, and ALL over time. We sort the time series top down in order of their cumulative return at the latest point in time series. As expected, the strategy ALL is best in class with a cumulated return value of 9.47. We observe that time series of ALL, BOO and RAF are flattening out since 2013. This may be the influence of better odds, as the bookmakers themselves started using machine learning algorithms which results in a more precise estimation of odds capturing the increasing amount of available information. These findings, especially the decrease in profit over time, confirm the weak efficiency hypothesis as pointed out in the following section.

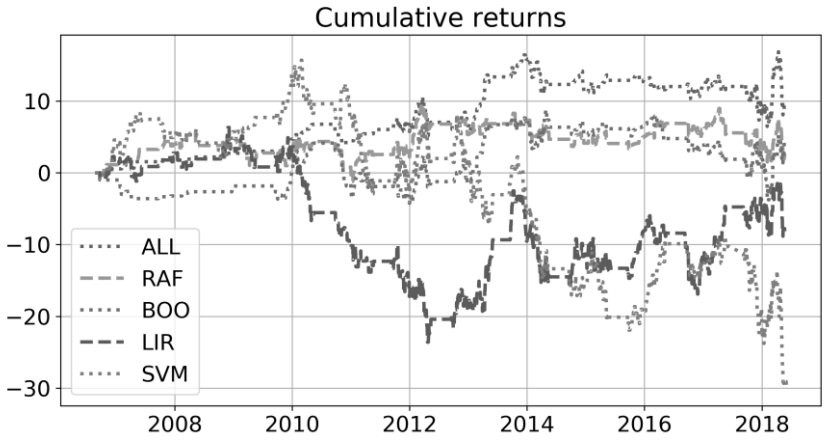


Figure 1. Cumulative returns of RAF, BOO, LIR, SVM, and ALL. from football seasons 2006/2007 to 2017/2018. Source: Stübinger et al. (2020).

3. Implications for the efficiency of the market

Market Efficiency states that all publicly available data is manifested in fair betting odds (up to a surcharge added to the odds by the bookmakers). This implies that no systematic outperformance should be possible.

Fama (1970) introduces the concept in the context of financial markets (weak form) which states abovementioned hypothesis. Lately, football betting has been getting more interest in research as there is a vast amount of publicly available data – both for players/teams statistics and various bookmakers odds. The latter should reflect all the information that can be derived from aforementioned data if the efficiency hypothesis holds true.

3.1. Inefficiencies in football odds

Angelini & Angelis (2018)[1] give an excellent introduction on challenging the efficiency hypothesis in the area of football. They analyze 11 major European football leagues covering 11 years of data and revealing inefficiencies in three of the leagues. However, the analysis is based on econometrical models, only, whereas Stübinger et al. (2020) cover the application of state-of-the-art machine learning technique with a strong emphasis on the prediction of the correct outcome of a game rather than modeling the mechanics between the influencing factors. Stübinger et al. (2020) point out diminishing profitability of

[1] See Angelini & Angelis (2018) and references therin.

systematic betting approaches over time, which can be explained by improved modeling techniques and the improved availability of processing power and data. In this section, we link the aforementioned results of Stübinger et al. (2020) to the efficiency hypothesis.

Efficiency in this context is always referring to betting odds including all available information. Thus, an information asymmetry would lead to systematic wins on average when used to identify 'weak' betting odds (of either the bookmaker or the betting person). On the other hand, when applying betting strategies that use only limited amount of information (compared to the odds), one can expect systematic losses over time. This is in line with the results of Stübinger et al (2020), where strategies that use no or little information (random-betting or favoring the home team) never gain any positive returns. However, using todays computation power and data, one can generate an historical information advantage in times where the odds did not reflect that knowledge which results in positive returns for earlier periods.

In older times, the process of reflecting believes about the outcome of a game using betting odds was in the hand of individual bookmaker companies. As the amount of information kept on growing, providers of betting-odds-as-a-service[2] started entering the market who offer real-time information before and during sport events leveraging enormous data collection tools and modelling power. Also, by providing the major bookmaker companies with similar betting odds, arbitrage effects are no longer available.

3.2. Efficiency and Machine Learning

The results of Stübinger et al. (2020) state that it would have been highly profitable to use machine learning based betting strategies in early 2000s. However, in those times neither the data nor the computation power has been available as easily as it is of today. By choosing the data source and the applied ML algorithms constant over time, we observe diminishing returns when using the predictions of a football game to bet against the bookmakers. We think that this is strongly related to the increased ability of odd providers to collect and process data together with more sophisticated modeling techniques. Whereas in older times a single, well informed person with knowledge and experience of analyzing games in his favorite league could have had an overview on all available data by screening the newspaper or relevant internet articles, nowadays in times of social media[3] and openly accessible datasets[4] this can only be achieved using algorithms and cloud computing. The information lead of bookmakers that is reflected by more accurate odds makes it impossible

[2] Such as https://www.betradar.com/ or https://txodds.net/.

3 E.g, Schumaker et al (2016) using Twitter data.

[4] https://datahub.io/collections/football

for individuals to beat the market in the long run, which is a strong confirmation for the market efficiency hypothesis.

3.3. Bookmakers perspective

Naturally, the view of the bookmakers plays an important role in the context of football betting. Provider determine the offered odds on the basis of a two-stage procedure. First, they estimate the probabilities of the outcomes home win, draw, and away win. The methods applied here vary between naive baseline approaches and highly complex machine learning models. The sum of the three probabilities is always 1. The fair betting odds would be the inverse of the respective probabilities. Second, bookmakers usually diminish the fair odd value by their risk premium, i.e., a certain amount to cover their costs in the long run. Consequently, we can calculate this key figure by determining the difference between the sum of the inverse of the actually offered betting odds and 1. A higher risk premium is tantamount to a more inefficient market environment as the bookmakers need to hedge their earnings due to the increased uncertainty of the outcome of a game. Figure 1 presents the average risk premium of the online bookmaker Bet365, one of the leading betting providers with around 23 million customers, from 2006 to 2018. We observe a decrease of the risk premium from around 10 percent in 2006 to around 5 percent in 2018. This fact is not surprising since more and more betting providers are entering the market and the information available is increasing. This picture is very similar to the increasing market efficiency as we know it very well from the stock market environment.

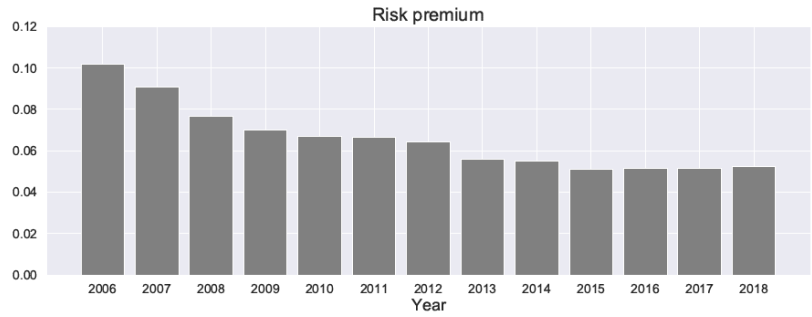


Figure 2. Average risk premium of the bookmaker Bet365 from 2006 to 2018.

Additionally, we want to provide some insights on the distribution of the risk premium across different type of football teams. Figure 3 displays the relation of average points and average risk premium for the teams of the Primera Division from 2006 to 2018. Top teams like Real Madrid and Barcelona achieves around 2.3 points per match, bad teams like Granada and Gijon around 1 point per match. We observe an average risk premium of approximately 6.2 percent for the top teams as well as the bad teams. Teams of average

quality in terms of long-run performance like Zaragoza and Mallorca possess a higher risk premium. We may conclude that is much more difficult to predict the outcome of matches involving teams of average quality.

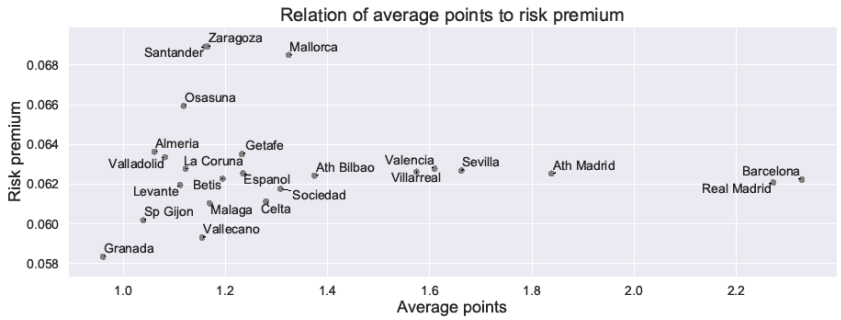


Figure 3. Relation of average points and average risk premium for the teams of the Primera Division from 2006 to 2018.

4. Outlook

In this paper we revisited the results of Stübinger et al (2020) from the perspective of the market efficiency hypothesis. We come to the conclusion, that diminishing returns over time placing bets based on signals generated by machine learning algorithms confirm that all available information is reflected in fair betting odds in recent times. The fact that using available information cannot systematically outperform the bookmakers confirms the weak efficiency hypothesis by Fama (1970) in recent times.

References

Angelini, G., & Angelis, L., (2019). *Efficiency of online football betting markets*. International Journal of Forecasting, 35(22), 712-721.

Fama, E. (1970). *Efficient Capital Markets: A review of theory and empirical work*. The Journal of Finance, 25(2), 383-417.

Schumaker, R. P., Jarmoszko , A. T., & Labedz, C. S. (2016). *Predicting wins and spread in the Premier League using a sentiment analysis of Twitter*. Decision Support Systems, 88, 76-84,

Stübinger, J., & Knoll, J. (2018). *Beat the Bookmaker – Winning football bets with machine learning (best application paper)*. In International Conference on Innovative Techniques and Applications of Artificial Intelligence (pp. 219-233). Springer, Cham.

Stübinger, J., Mangold, B., & Knoll, J. (2020). *Machine learning in football betting: Prediction of match results based on player characteristics*. Applied Sciences, 10(1), 46.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11621

Sentiment Analysis of Twitter in Tourism Destinations

Carmen Pérez Cabañero, Enrique Bigné, Carla Ruiz, Antonio Carlos Cuenca
Department of Marketing, University of Valencia, Spain.

Abstract

Given the importance of electronic word of mouth (eWOM), this paper analyses the content of messages generated by users related to a tourist destination and shared through Twitter. We propose three research questions regarding eWOM behaviour in Twitter focused on the expertise of the reviewer, sentiment analysis of a tweet and its content. In order to address those research questions we carry out text mining analysis by retrieving existing information on Twitter (over 1500 tweets) regarding to Venice as a tourist destination.

Keywords: *Twitter; eWOM; tourism, sentiment analysis.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Social media have deeply changed the way users search tourism information and share their travel experience, emotions and experiential moments. Electronic word of mouth (eWOM) refers to "any statement made by potential, actual or former consumers about a product, service or company, which is available to a multitude of people and institutions via Internet" (Hennig-Thurau et al., 2004). Despite tremendous attention to eWOM in tourism research, destinations have attracted only 10% of the papers published from 2009 to 2016 in major tourism and hospitality journals (Sotiriadis, 2017).

2. Goals

This paper analyses the content of messages generated by users (User Generated Content, UGC) related to a tourist destination and shared through Twitter. More specifically we focus on the sentiment of tweets. Sentiment analysis refers to the subjective value of the content of the online comments which is typically expressed as positive or negative (Alaei, Becken and Stantic, 2019).

Yoo and Gretzel (2011) have found that United States travelers as creators of UGC are mostly motivated by altruistic and hedonic benefits. In contrast, self-centred motivations include possibilities for gaining recognition, increasing social ties and augmenting one's self-esteem, among others (Gretzel and Yoo, 2008). Based on self-centered motivations the end goal of a tweet is to be influential. Research on the influence or centrality has recently emerged through different measures (see for a review Riquelme and González-Cantergiani, 2016). We wonder whether experienced users post more positive or negative comments. Therefore we posit the following research question: RQ1. Does the expertise of the reviewer influence on the sentiment of his/her tweets?.

Stieglitz and Dag-Xuan (2013) argue that the expression of emotions in social media–based textual content may also lead to more attention and arousal, which in turn may positively affect information sharing behaviour. According to their empirical results, twitter messages that feature a high degree of emotionality tend to trigger more retweets. In order to confirm this outcome, we propose: RQ2. Does the sentiment of a tweet impact on eWOM behaviour in Twitter?

Previous research suggests that there is only a small percentage of tweets that participants retweet. Tweets containing links are rated as being significantly more interesting than tweets without links, but hashtags make no difference in terms of perceived interest (Counts and Fisher, 2011). We wonder about the impact of several components of the tweet on eWOM behavior. Therefore, we propose the following research question: RQ3. Which of the

information cues of a tweet (images, links, hashtags, bookmarks) can best predict eWOM behavior in Twitter?

3. Methodology

We carry out text mining analysis by retrieving existing information on Twitter (over 14,000 tweets) regarding Venice as a tourist destination. The selection of the comments under analysis is based on the mention ten selected keywords for analysis: Tourism, holiday in Venice, travel Venice, getaway to Venice, booking Venice, Venice port, weather in Venice, venice hotels, flights to Venice and, to see in Venice. We asses expertise of a Twitter user derived from the number of years using Twitter, number of tweets made, and number of subscribed lists. eWOM behaviour is measured based on the number of retweets, the user's reachness and the tweet's effective reachness. Sentiment analysis was made using Meaning Cloud. We obtained an ordinal variable for classifying the content of the tweets from 1 very negative to 5 very positive. Statistical analysis was carried out using IBM SPSS 26.

4. Results

The sample is made up of 14,338 tweets collected during July and August of 2016. There are 6,352 original tweets, 311 are responses and 7,675 are retweets. Regarding the demographics of users, 4,500 were posted by women and 2,082 men, all of them are originally posted in English but the origin and the age of users is unknown.

Regarding the first research question relating the expertise of the user and the sentiment of tweets, we will select only original tweets (reponses and retweets will not be considered). A regression analysis reveals the different impact of the number of years using Twitter, the number of tweets made and the number of lists subscribed (see table1).

Table 1. Regression analysis on Sentiment.

	Unstand. Beta	**St.coef. Beta**	**t**	**Sig**	**Tolerance**	**FIV**
Constant	4.187		169.653	0.000***		
Years	-0.054	-0.152	-8.783	0.000***	0.970	1.031
Number of lists	2.767E-6	0.110	5.938	0.000***	0.838	1.193
Number of tweets	-4.791E-7	-0.065	-3.444	0.001***	0.815	1.226

R= 0.190; R2= 0.036; F = 41.440 (Sig= 0.000); Durbin-Watson= 1.605; ***=p<0.01
Dependent variable: Sentiment

Therefore, the more years the user is using Twitter, the more negative the messages. This variable is the most influential of the analysis (its standadrdized Beta of -0.152 is the highest). Addtionally, the more lists the user is susbscribed, the more positive the tweets of that user. This positive relationship is illustrated by the standadrdized Beta of 0.110. Finally, the users with a lot of tweets posted have a tendency to post negative ones as shown by its standadrdized Beta of -0.065 which is quite low but significant.

Regarding the second research question about the impact of the sentiment of tweets on eWOM, we carried out a ANOVA with the subsample of the original tweets. The variable sentiment is the factor in our analysis. We considered several aspects of eWOM such as the number of retweets achieved, the user's reachness and the tweet's effective reachness. User's reachness is measured adding the user's followers plus the followers of those retweeting and those answering the tweet. The tweet's effective reachness is measured adding the number of answers, the number of tweets with mentions and the number of retweets. The data offered heterocedascidity of variances so we report the Welch statistic as a Robust test of equality of means and the Games-Howell post hoc analysis. These results are shown in Table 2 though we only show the cells with significant relationships.

In order to understand the results obtained out of the comparison of means above, we include the Table 3 which contains the main descirptives of this analysis.

The results in Table 2 show that the are different means for the number of retweets reached between neutral and positive reviews. According to results in Table 3, Neutral tweets only make an average of 0.24 retweets while Positive tweets make an average of 2.57 retweets. In general terms, the more positive the tweet the higher the number of retweets it achieves. There are also different means for the user's reachness regarding Very negative and Positive tweets. According to Table 3, Very negative tweets achieve an average of 2,259.72 followers while Positive tweets achieve an average of 19,787.69 followers. Thus, the reachness of Very negative tweets is the least of all tweets. With respect to the effective reachness of the tweets, there are statistical differences between the mean of Very negative tweets and the means of Negative, Positive and Very positive tweets. According to Table 3, the mean of Very negative tweets is 718.35, which is the lowest score in that calculation. Concluding this analysis, we can state that Very negative and negative tweets have less impact on eWOM behavior.

Table 2: Welch statistics and Games-Howell post hoc analysis.

Dependent variable			Mean difference	Std. error	Sig.
Number of retweets	3 Neutral	1 Very negative	-0.0008	0.180	1.000
		2 Negative	-1.722	0.958	0.377
		4 Positive	-2.332	0.852	0.049*
		5 Very positive	-157.247	141.224	0.799
User's reachness	1 Very negative	2 Negative	-40441.541	30111.98	0.664
		3 Neutral	-2982.310	1902.231	0.521
		4 Positive	-17527.97	3151.534	0.000***
		5 Very positive	-61102.612	29754.87	0.242
Tweet's effective reachness	1 Very Negative	2 Negative	-4940.020	1790.307	0.048*
		3 Neutral	-1914.375	1557.351	0.734
		4 Positive	-6071.102	1021.187	0.000***
		5 Very positive	-5878.971	1307.420	0.000***

***=p<0.001; *=p<0.05

Number of retweets: Welch statistic= 2.962; df1=4; df2=322.452; Sig.=0.020

User's reachness: Welch statistic= 9.341; df1=4; df2=619.169; Sig.=0.000

Tweet's effective reachness: Welch statistic= 12.846; df1=4; df2=483.425; Sig.=0.000

Table 3. Descriptives.

		N	Mean	Std. Deviation
Number of retweets	1 Very negative	37	0.243	1.011
	2 Negative	347	1.965	17.811
	3 Neutral	99	0.242	0.701
	4 Positive	1997	2.574	37.945
	5 Very positive	869	157.490	4163.130
User's reachness	1 Very negative	37	2259.729	5041.584
	2 Negative	347	42701.270	560711.589
	3 Neutral	99	5242.040	17035.862
	4 Positive	1997	19787.699	135877.437
	5 Very positive	869	63362.341	876797.807
Tweet's effective reachness	1 Very negative	37	718.351	2804.553
	2 Negative	347	5658.371	32224.824
	3 Neutral	99	2632.727	14800.793
	4 Positive	1997	6789.453	40718.484
	5 Very positive	869	6597.322	36065.084

Finally we examine the information cues of a tweet (images, mentions, and hashtags) which impact on eWOM behavior in Twitter. These information cues are nominal variables (whether it is included or not in the tweet) so we produce several t-test where the dependent variable refers to eWOM behavior, that is the number or retweets, the user's reachness and the tweet's effective reachness. As eWOM behavior is under study, we select original tweets in our sample. The descriptives about the tweets containing hashtags, mentions and images are shown in next Table 4.

Table 4. Descriptives according to cues.

	Hashtag	**N**	**Mean**	**Std. deviation**
Number of retweets	No	2818	46.45	2301.98
	Yes	3534	4.40	195.51
User's reachness	No	2818	32636.43	540239.86
	Yes	3534	26616.84	178098.94
Tweet's eff. reachness	No	2818	6218.54	144704.06
	Yes	3534	15209.52	57711.21
	Mention	**N**	**Mean**	**Std. deviation**
Number of retweets	No	5613	22.81	1630.93
	Yes	739	24.93	432.14
User's reachness	No	5613	24198.18	379868.61
	Yes	739	67941.79	408622.99
Tweet's eff. reachness	No	5613	11234.27	111014.91
	Yes	739	11118.18	48214.26
	Picture	**N**	**Mean**	**Std. deviation**
Number of retweets	No	406	15.14	91.74
	Yes	5946	23.60	1591.71
User's reachness	No	406	37267.21	256520.91
	Yes	5946	28742.50	390727.29
Tweet's eff. reachness	No	406	257.55	1561.60
	Yes	5946	11969.35	109149.54

The results regarding the t-test analyses are shown in the Table 5 below. Firstly, the assumption of equal variances is checked by means of the F statistic. Thus, in case the significance associated to the F statistic is higher than 0.05 we report the corresponding t statistic when equal variances are assumed. Alternatively, in case the significance associated to the F statistic is lower than 0.05 we report the corresponding t statistic when equal variances are not assumed.

Table 5. Results of t-test analyses.

	Hashtag	**F**	**Sig.**	**t**	**df**	**Sig.**
Number of retweets	Equal variances not assumed	4.612	0.032*	0.967	2849.42	0.334
User's reachness	Equal variances assumed	3.498	0.061	0.621	6350	0.534
Tweet's eff. reachness	Equal variances not assumed	27.472	0.000***	-3.107	3531.59	0.002**
	Mention	**F**	**Sig.**	**t**	**df**	**Sig.**
Number of retweets	Equal variances assumed	0.001	0.973	-0.035	6350	0.972
User's reachness	Equal variances not assumed	22.022	0.000***	-2.757	913.93	0.006**
Tweet's eff. reachness	Equal variances assumed	0.100	0.752	0.028	6350	0.978
	Picture	**F**	**Sig.**	**t**	**df**	**Sig.**
Number of retweets	Equal variances assumed	0.049	0.825	-0.107	6350	0.915
User's reachness	Equal variances assumed	0.973	0.324	0.433	6350	0.665
Tweet's eff. reachness	Equal variances not assumed	14.943	0.000***	-8.262	5979.90	0.000***

***=p<0.01; **=p<0.005; *=p<0.05

The results in Table 4 and 5 show the influence that several cues included in the tweets have on eWOM behavior. Thus, we can conclude that tweets having hashtags have a higher effective reachness (t= -3.107, sig.=0.002). Tweets having mentions of other Twitter users increase user's reachness (t= -2.757, sig.=0.006). Finally, tweets having pictures increase their effective reachness (t= -8.262, sig.=0.000).

In conclusion, this paper analysed the content of a collection of tweets related to Venice as a tourist destination and applied sentiment analysis. Several aspects related to the user's expertise impact on the sentiment of tweets posted like the number of years the user is in Twitter and the tweets posted overall. Sentiment of tweets also impact on eWOM behaviour.

The more positive the tweet the higher the number of retweets it achieves while negative tweets have less impact on eWOM behavior. Regarding the content analysis of the tweet, having hashtags and pictures impact on the tweet's effective reachness while having mentions impact on the user's reachness. Management destination organisations could benefit from current results to make a higher impact of their tweets, for example including hashtags and pictures to improve the tweet's effective reachness and including mentions to gather potential new followers. Further research can enlarge the sample of tweets under study and compare different tourist attractions like free attractions (for instance a park, a main square or a cathedral) and paid-for attractions (like museums, castles and private buildings).

Acknowledgements

Authors acknowledge financial support of research project UV-INV_AE19-1212255.

References

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. *Journal of Travel Research*, 58(2), 175-191.

Counts, S., & Fisher, K. (2011). Taking It All In? Visual Attention in Microblog Consumption. *ICWSM*, 11, 97-104.

Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. *Information and communication technologies in tourism 2008*, 35-46.

Hennig-Thurau T., Gwinner K., Walsh G., & Gremler d. (2004). Electronic Word-of-Moud Vía Consumer –Opinion Platforms: What Motivates Consumer to Articulate Themselves on the internet. *Journal of Interactive Marketing*, 18(1), 38-52.

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. *Information Processing & Management*, 52(5), 949-975.

Sotiriadis, M. D. (2017). Sharing tourism experiences in social media. *International Journal of Contemporary Hospitality Management,* 29(1), 179-225.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4), 217-248.

Yoo, K. H., & Gretzel, U. (2011). Influence of personality on travel-related consumergenerated media creation. *Computers in Human Behavior*, 27(2), 609-621.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11622

Google Trends Topic-Based Uncertainty: A Multi-National Approach

Florian Schütze
Department of Socioeconomics, University of Hamburg, Germany.

Abstract

Several studies have shown that uncertainty among economic actors influences business cycle dynamics. This paper uses Google Trends topic queries to construct an uncertainty proxy that can be applied to every country where Google is active. Using a VAR approach, this paper demonstrates that the obtained impulse-response functions of main economic indicators to a one-standard deviation shock to the constructed indicator, are similar to those from an already-existing uncertainty proxy, the EPU. This is true for the G7 countries and Russia. On average, the uncertainty indicator constructed for this paper leads to more statistically significant responses than does the EPU. Thus, this paper shows that Google Trends is a helpful tool for obtaining timely information about uncertainty among economic actors. The main improvement in this uncertainty proxy is in its language independence. Existing uncertainty-measurement approaches, in contrast, rely on certain keywords that often vary across countries.

Keywords: *Google Trends; Uncertainty; Business Cycle Dynamics; VAR.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

A variety of authors have demonstrated that uncertainty among economic actors has an influence on business cycle dynamics (Federal Open Market Committee, 2009; Gilchrist et al., 2014; Fernández-Villaverde et al., 2015; Jones & Olson, 2015; Segal et al., 2015; Leduc & Liu, 2016; Shoag & Veuger, 2016; Moore, 2017; Mumtaz & Theodoridis, 2018). Uncertainty is a latent variable and cannot be directly observed, so there are three indirect measurement methods: measuring volatility in various metrics, such as forecast errors (Jurado et al., 2015) and stocks (Bloom, 2009); measuring dispersion among forecasters or in business surveys about economic tendencies (Bachmann et al., 2013); and measuring the occurrence of keywords in certain mediums, such as newspapers, as it is done by Baker et al. (2016) for the Economic Policy Uncertainty index (EPU).

In this paper, an uncertainty indicator based on the third method is constructed using Google Trends search topics. This indicator can be applied directly to all countries that use Google. The advantages of this approach over the EPU are threefold. First, it does not rely on journalists writing about their own perceptions of economic uncertainty—it measures it directly from economic actors. Second, it is not only limited to uncertainty over economic policy—it covers overall economic uncertainty. Third, it can be conducted daily in real time.

The results of this paper show that shocks to the constructed uncertainty proxy in the G7 countries and Russia lead to similar VAR impulse-response functions compared to shocks to the EPU. On average, the uncertainty indicator constructed for this paper leads to more statistically significant responses than does the EPU. The difference in significant responses between them is also, in turn, statistically significant.

Previously, the use of Google Trends to construct an uncertainty indicator has been done primarily with keywords (e.g., Castelnuovo & Tran, 2017; Bontempi et al., 2018; Donadelli & Gerotto, 2019). The works that have employed this method all show that their proxy influences business cycle dynamics while providing information quickly, relative to other uncertainty proxies. The problem with this approach, however, remains—for every different language, there must be a unique uncertainty-related word set.

Fortunately, Google Trends offers the ability to search for topics rather than keywords. Thus, Google presents a serious advantage; it uses machine-learning techniques to identify the underlying topics in search queries. With this option, extending the uncertainty proxy to every Google-using country is fairly straightforward, provided that the relevant topics are identified. Until now, this approach had only been used by Kupfer and Zorn (2019) with ten topics and four categories; no selection criteria for the topics or categories were provided. Nonetheless, they found a statistically significant influence on stock market volatility.

This paper is structured as follows. The following section provides insight into the theoretical background on uncertainty's importance matters; it also explains why the Google Trends approach is suited for this task. The third section offers an overview of the empirical procedure used for this paper. The fourth section presents the results of the VAR analysis. Finally, the fifth section concludes the paper.

2. Theoretical Background

The idea that Google Trends can accurately measure economic uncertainty relies on two assumptions. First, uncertain actors aim to reduce their uncertainty by gathering information on a certain subject. Second, actors using the Internet to gather information do so primarily using search engines, among which Google is the most significant. With these two assumptions, one can conclude that higher search requests reflect higher uncertainty.

In terms of the theoretical background, the reason uncertainty has a negative influence on the economy is twofold (Bloom, 2014). On the one hand, consumers and investors may postpone actions when uncertainty is high. This is called the "wait-and-see" effect when said postponement is accompanied by an overshooting of consumption and investment after uncertainty levels return to normal. On the other hand, uncertainty may be interpreted as constituting a high risk[1] and, therefore, overall risk premiums rise, leading to an increase in the cost of investments.

A VAR is used to estimate impulse-response functions of bonds, shares, and industrial production after a shock of uncertainty. The response of bond yields to a shock of uncertainty can be positive or negative, depending on the attitude of economic actors toward the government. A positive response suggests that people demand a risk premium due to high uncertainty. A negative response suggests that people have high trust in their governments and buy these bonds, leading to a decrease in the yield. It is assumed that elevated uncertainty negatively impacts stock market returns because people sell their shares with a "wait-and-see" attitude. It is also assumed that high uncertainty negatively impacts industrial production.

3. Empirical Approach

3.1. Data

This paper used data from Google Trends for its empirical analysis. The data was obtained using the "gtrendsR" package for the program "R." Data on stock market returns, long-term

[1] Here, it is important to distinguish uncertainty from risk. See Knight (1921) for definitions of risk and uncertainty.

government bond yields, and industrial production for the G7 countries and Russia is also used (Data source: OECD, 2019c; OECD,2019b; OECD, 2019a). All data was integrated of order one to obtain stationary time-series and seasonal-adjusted data using "X-13ARIMA-SEATS."

For stock market returns, this paper uses the major stock index of each country measured in monthly growth rates. For changes to long-term government bond yields, bonds with a maturity of ten years are used. Industrial production is represented by annualized monthly growth rates of said industrial production.

The period considered for the G7 countries is from 02/2004 to 02/2019; for Russia, the period is from 02/2004 until 06/2018, as the OECD only has long-term Russian government bond data up to 06/2018.

3.2. Construction Approach

To identify economic uncertainty-related topics, this paper used 184 Google Trends queries stemming from Bontempi et al. (2018), which are based on category-specific policy keywords by Baker et al. (2016). The underlying topics were identified by inserting the keywords into Google Trends and saving the first returned topic. After deleting recurring ones, the process ultimately returned 156 topics. Additionally, the four news categories suggested by Kupfer and Zorn (2019) were also included.

These 160 potential search queries were then downloaded for all countries being considered. Nineteen queries that returned no data for at least one considered country were excluded, leaving us with 141 queries.

In order to keep only queries that are highly informative, the data for the US and Canada is used as a training set. As the initial keywords from Baker et al. (2016) were crafted for the construction of the EPU from US newspapers, the topic selection needed to be done with US data. However, since this could lead to very US-specific topics, Canadian data was included to mitigate this effect.

Ultimately, 141 US and Canadian VARs were estimated in the following order: a single query out of all 141; long-term government bond yields; stock market returns; industrial production. Since we are interested in the contemporaneous effect of the resulting impulse-response functions, the transmission is assumed to follow this order. The lag length of the VAR was chosen by AIC with a maximum lag of 12 months.

For construction, only topics and categories with at least three statistically significant responses of the US and Canadian VAR are kept. The corresponding impulse-response functions stem from how stocks, bonds and industrial production respond to a one-standard deviation shock in a Google Trends query.

The in the paragraph above mentioned procedure resulted in 13 queries: twelve topics and one news category. Only one query, namely the news category, is identical to the 14 queries used by Kupfer and Zorn (2019). To construct the Google Trends Topic Uncertainty indicator (GTTU) for each country, the remaining queries were aggregated in accordance with the suggestion made by Bontempi et al. (2018) to obtain uncertainty indicators for each country.

A VAR was constructed for the G7 countries and Russia in the following order: the country-specific aggregated GTTU; country-specific long-term government bonds; country-specific stock market returns; country-specific industrial production. Once again, the lag length of the VAR was chosen by AIC with a maximum lag of 12 months.

4. Country-Specific Results

Figure 1 is a comparison between this paper's US GTTU (GTTU_US) and the US EPU (EPU_US) from Baker et al. (2016). The GTTU shows a sharp spike in uncertainty around the 2008/2009 financial crisis and two smaller spikes in 2016 and 2018. The EPU shows spikes around the 2008/2009 financial crisis as well as in 2012 and 2016, which could reflect political uncertainty regarding US elections. It also shows a clear spike in 2018, which is likely attributable to the US-China "trade war."

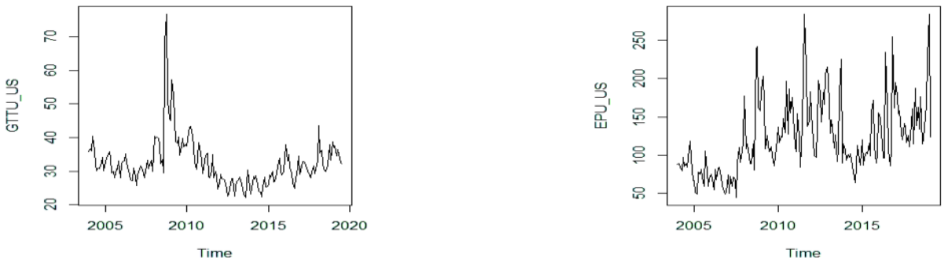
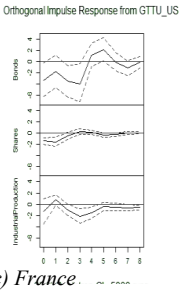
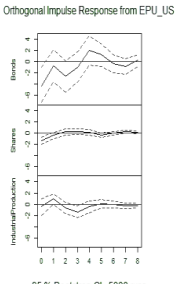


Figure 1. Uncertainty over time in the USA (GTTU_US vs. EPU_US).

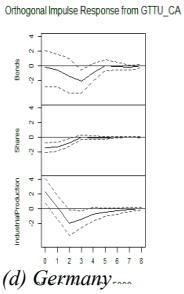
Figure 2 presents the impulse responses to a one-standard deviation shock to uncertainty for all of the considered countries. Once more, it displays this paper's GTTU and the EPU.

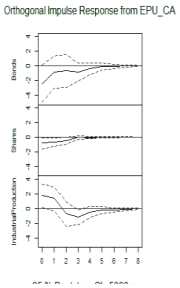
(a) United States



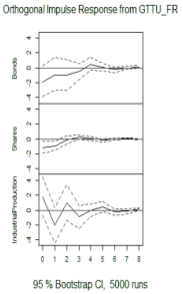


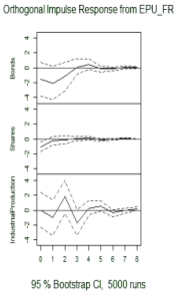
(b) Canada



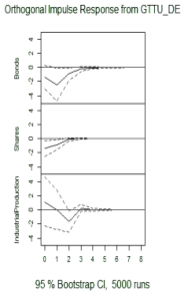


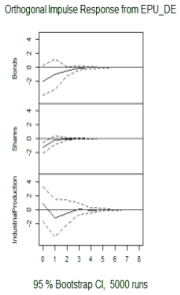
(c) France



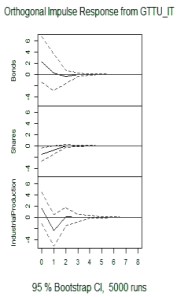


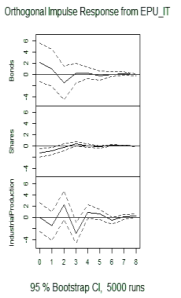
(d) Germany



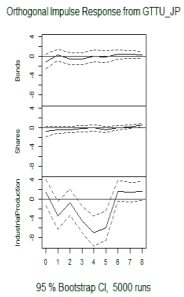


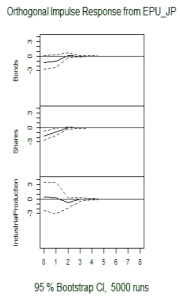
(e) Italy



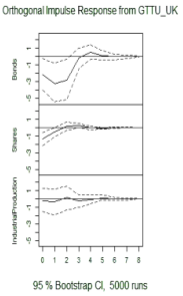


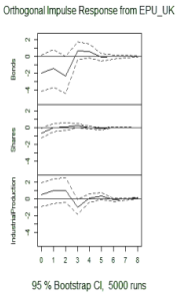
(f) Japan



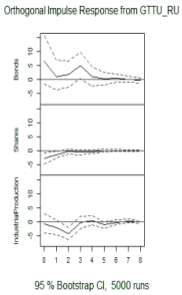


(g) United Kingdom





(h) Russia



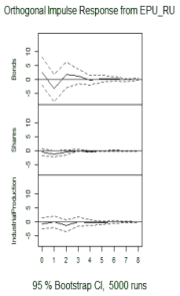


Figure 2. Impulse-response functions to a one-standard deviation shock to uncertainty.

For the US, the responses are similar for both uncertainty indicators, though the GTTU offers more statistically significant responses (ssr) than does the EPU (a total of 13 GTTU ssr compared to 4 EPU ssr). The results for Canada are very similar, the only difference being that shares and industrial production show a stronger significant response from a GTTU shock compared to an EPU shock (7 GTTU srr, 4 EPU ssr). The two uncertainty measurements show similar results for France (3 GTTU ssr, 1 EPU ssr). German results indicate that, for bonds, only the GTTU shows a significant response. For shares, the significant response is one month longer for the GTTU. Industrial production shows a significant response in the second month for the GTT (5 GTT ssr, 1 EPU ssr).

For Italy, the statistically significant responses for shares are one month longer for the EPU. The other two responses show the same directions for both proxies, though they are only for the EPU and industrial production significant(2 GTTU ssr, 5 EPU ssr). The Japanese case is an outlier, as it shows a very strong response from industrial production to a GTTU shock (5 GTTU ssr, 2 EPU ssr). For the UK, the main difference in the impulse response functions lies in the response of bonds to a GTTU shock, which is significant for up to two months (5 GTTU ssr, 3 EPU ssr). For Russia, there is a significant response from bonds, shares, and industrial production to a GTTU shock; in contrast, an EPU shock results only in a significant response from shares (4 GTTU ssr, 1 EPU ssr).

In conclusion, similar directions of impulse responses to shocks to the EPU and the GTT suggest that they are both measuring the same hidden uncertainty. While stocks and bonds react alongside uncertainty shocks, industrial production generally reacts after about two months. For all countries considered, the response's direction is in line with the theoretical prediction. Interestingly, the reaction of government bonds is negative in all but Italy and Russia, suggesting that economic actors demand a risk premium amid high uncertainty in both countries. Importantly, the GTTU outperformed the EPU in terms of statistically significant responses in all of the considered countries except Italy. A one-standard deviation shock of the GTTU uncertainty proxy results in, on average, 5.5 statistically significant responses. The same shock to the EPU results in an average of just 2.63 statistically significant responses. Therefore, the GTTU garners, on average, 109% more statistically significant responses than does the EPU, and this difference is significant at the 5% level. The difference for each impulse-response category is significant at least at the 10% level. Furthermore, the GTTU shows a significant response of share returns to uncertainty shocks for all considered countries. For five countries, it shows a significant response of bond yields and industrial production. In contrast, the Google Trends topic-based uncertainty proxy constructed by Kupfer and Zorn (2019) yielded far fewer significant responses (shares: US, UK, and France; bonds: US and UK; industrial production: Canada).

5. Conclusion

This paper detailed the construction of an uncertainty indicator based on Google Trends topics. As shown, the resultant GTTU uncertainty proxy resembles the EPU but performs better—in terms of significant impulse responses—in seven of the eight countries considered.

While Kupfer and Zorn (2019) had already produced a Google Trends topic-based uncertainty-measurement technique, they presented no procedure for choosing the topics. In this paper, the selection was based on the category-specific policy terms set by Baker et al. (2016). This resulted in 141 unique topics, which were further reduced by using the US and Canadian economic data collectively as a learning set in order to only keep topics relevant to the business cycle. Ultimately, 13 highly informative topics came to make up the Google Trends Topic Uncertainty indicator for each country.

With this uncertainty measurement, it is possible to construct a timely indicator for every country where Google is used, regardless of language. For example, when using English keywords for Japanese Google requests, only English-writing users would be considered. Furthermore, ambiguous keywords that may be present in some languages but not in others are not a concern, as all relevant keywords should be included in the topics.

The GTTU crafted for this paper exhibits behavior similar to that of the EPU from Baker et al. (2016). Moreover, it produces more statistically significant impulse responses than does the EPU, and this difference is statistically significant.

With the options provided by Google Trends, this indicator can operate on a daily basis. In addition to countries, it could apply to subregions like US states. There is a multitude of potential applications for this indicator that future research should consider. For example, researchers could use it to examine how citizens view their governments. As shown in the last chapter, reaction directions from government bond yields vary by country; these differences could be interpreted as differences in public attitude toward the state.

References

Bachmann, R., Elstner, S., & Sims, E. R. (2013). Uncertainty and Economic Activity: Evidence from Business Survey Data. *American Economic Journal: Macroeconomics*, *5*(2), 217–249.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636.

Bloom, N. (2009). The Impact of Uncertainty Shocks. *Econometrica*, *77*(3), 623–685.

Bloom, N. (2014). Fluctuations in Uncertainty. *Journal of Economic Perspectives*, *28*(2), 153–176.

Bontempi, M. E., Frigeri, M., Golinelli, R., & Squadrani, M. (2018). Uncertainty, Perception and Internet, Working Papers wp1134, Dipartimento Scienze Economiche, Universita' di Bologna.

Castelnuovo, E. & Tran, T. D. (2017). Google It Up! A Google Trends-based Uncertainty index for the United States and Australia. *Economics Letters*, *161*, 149–153.

Donadelli, M. & Gerotto, L. (2019). Non-macro-based Google searches, uncertainty, and real economic activity. *Research in International Business and Finance*, *48*, 111–142.

Federal Open Market Committee (2009). Minutes of the Federal Open Market Committee: December 15-16, 2009.

Fernández-Villaverde, J., Guerrón-Quintana, P., Kuester, K., & Rubio-Ramírez, J. (2015). Fiscal Volatility Shocks and Economic Activity. *American Economic Review*, *105*(11), 3352–3384.

Gilchrist, S., Sim, J. W., & Zakrajsek, E. (2014). Uncertainty, Financial Frictions, and Investment Dynamics: Working Paper.

Jones, P. M. & Olson, E. (2015). The International Effects of US Uncertainty. *International Journal of Finance & Economics*, *20*(3), 242–252.

Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring uncertainty. *American Economic Review*, *105*(3), 1177–1216.

Knight, F. H. (1921). *Risk, uncertainty and profit*. Boston: Houghton Mifflin.

Kupfer, A. & Zorn, J. (2019). A Language-Independent Measurement of Economic Policy Uncertainty in Eastern European Countries. *Emerging Markets Finance and Trade*, *4*(1), 1–15.

Leduc, S. & Liu, Z. (2016). Uncertainty shocks are aggregate demand shocks. *Journal of Monetary Economics*, *82*, 20–35.

Moore, A. (2017). Measuring Economic Uncertainty and Its Effects. *Economic Record*, *93*(303), 550–575.

Mumtaz, H. & Theodoridis, K. (2018). The Changing Transmission of Uncertainty Shocks in the U.S. *Journal of Business & Economic Statistics*, *36*(2), 239–252.

OECD (2019a). *Industrial production (indicator)*. Main Economic Indicators (database): DOI: 10.1787/data00052-en (Accessed on 18 July 2019).

OECD (2019b). *Long-Term Government Bond Yields (10-years)*. Main Economic Indicators (database): DOI: 10.1787/data-00052-en (Accessed on 18 July 2019).

OECD (2019c). *Share Prices (Indicator)*. DOI: 10.1787/6ad82f42-en (Accessed on 18 July 2019).

Segal, G., Shaliastovich, I., & Yaron, A. (2015). Good and bad uncertainty: Macroeconomic and financial market implications. *Journal of Financial Economics*, *117*(2), 369–397.

Shoag, D. & Veuger, S. (2016). Uncertainty and the geography of the great recession. *Journal of Monetary Economics*, *84*, 84–93.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11627

Bridging internet and cultural heritage through a digital marketing funnel: An exploratory approach

Călin Vegheș
Department of Marketing, Bucharest University of Economic Studies, Romania.

Abstract

Digital marketing and cultural heritage: what may have in common two areas that seem so different? What may connect a dynamic, evolving and even catchy field to a rather static, outdated and quite boring one? Maybe a funnel. Actually, a marketing funnel. More precisely a digital marketing funnel aiming to support the capitalization of the cultural heritage by drawing attention, raising interest, stimulate desire and generate action related to the cultural heritage output – goods, services, brands, events, and activities – under of the forms of discovering, exploring, experiencing and enjoying this heritage. Using secondary data regarding the cultural heritage in the European Union, the paper investigates the connections between the usage of the internet for cultural heritage purposes, different forms of consumption of the cultural heritage, and main barriers limiting this consumption and illustrates that building and employment of a digital marketing funnel is indispensable in the capitalization of the cultural heritage.

Keywords: *digital marketing; cultural heritage; marketing funnel.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

The definition issued by the ICOMOS International Cultural Tourism Committee (2002) sees the cultural heritage as the expressions of the ways of living developed by a community and passed on from generation to generation, including customs, practices, places, objects, artistic expression and values taking tangible (places of human habitation, villages, towns, and cities, buildings, structures, artworks, documents, handicrafts, musical instruments, furniture, clothing and items of personal decoration, religious, ritual and funerary objects, tools, machinery and equipment, and industrial systems) or intangible (all forms of traditional and popular or folk culture, the collective works originating in a given community and based on tradition – oral traditions, customs, languages, music, dance, rituals, festivals, traditional medicine and pharmacopeia, popular sports, food and the culinary arts and all kinds of special skill connected with the material aspects of culture) forms.

Researching the cultural heritage from a marketing perspective requires, besides this almost exhaustive and heritage experience-oriented vision, a more structured framework describing the cultural heritage market and related heritage consumption behavior. The European Commission (2017) has structured the cultural heritage market in libraries and archives; historical monuments and sites; museums and galleries; traditional events; traditional craft workplaces; cinema or film heritage festivals; traditional or classical performing arts events.

As the data of the Special Eurobarometer 466 reveals, the cultural heritage consumption of the European consumers is rather modest in spite of an overall context in which access to culture implying the consumption of various cultural goods and services by the public at large represents an opportunity to benefit from the cultural offer (Pasikowska-Schnass, 2017). Besides the demographical, economic, social or even… cultural reasons, the modest level of cultural heritage consumption can also be the result of the rare and less effective marketing campaigns conducted to convince the European consumer to discover, explore, experience and enjoy this heritage. What seems to be missing is a marketing funnel designed considering, on a hand, the stages of heritage cycle proposed by Thurley (2005) – understanding, valuing, caring for and, finally, enjoying, and, on the other hand, the classical stages of the AIDA model – attention, interest, desire, and action in connection to the cultural heritage.

In a context in which, on a hand, culture tends to become the fourth pillar of the sustainable development and culturally sustainable development adequately encompasses all the meanings of culture and all its complex interactions with the social, economic and environmental dimensions of human life (Sabatini, 2019), and, on the other hand, digital technologies are rapidly changing the environment by reducing information asymmetries between customers and sellers, and significantly changing the consumer behavior (Kannan and Li, 2017), designing a digital marketing funnel appears as the best way to promote and capitalize the cultural heritage consumption. Observing the increasing prevalence of digital

media and tools in marketing, Leeflang et al. (2014) have identified three major digital changes, such as the ability to interact with and/or serve customers in a new manner, increasing access to data and insights, and the ability to reach new customer segments that be considered in the design process. Going beyond digital customers' acquisition and retention, Eigenraam et al. (2018) have developed a taxonomy of customers' digital brand engagement practices to integrate ample research about such digital practices, and to standardize these digital practices across digital channels and platforms including five different types engagement practices (fun, learning, giving feedback, talk about and work for a brand) with corresponding tools. Integrating the digital component in the marketing funnel may improve the content and effectiveness of the marketing efforts conducted by the cultural heritage organizations addressing increasingly connected audiences in terms of facilitating discovery, exploration, experiencing and enjoying the cultural heritage.

2. Methodological notes

Data from the Special Eurobarometer 466 on Cultural Heritage (2017) have been considered in order to measure the ways in which European consumers have used the Internet for cultural heritage purposes, the frequency with which they attended cultural heritage-related activities and the main barriers encountered experiencing cultural heritage.

The variables of the research have been defined preserving the original format described in the Special Eurobarometer 466, as it follows:

1. Internet usage (IU), with the following sub-variables: IU0 – At least one cultural heritage related purpose; IU1 – Looking up general information related to cultural heritage, such as the accessibility, facilities and main features of a museum, historical monument, or traditional event in preparation for a visit or your holidays; IU2 – Buying or booking services for events or activities, such as tickets, guided tours, etc.; IU3 – Viewing cultural heritage-related content, such as the description of a work of art or historical monument during a visit, historical information about a traditional event you attend, etc.; IU4 - Creating or sharing cultural heritage-related content, such as a picture or a video of a work of art or historical monument, etc.; IU5 – Knowing more about a museum or a traditional festival, historical monuments, exhibition after a visit; and IU6 – Giving opinions of a cultural heritage site or activity (e.g. comments or scores on a review website).

2. Cultural heritage consumption forms (HC), with the following sub-variables: HC1 – Visited a library or archive (e.g. to consult manuscripts, documents, ancient maps, etc.); HC2 – Visited a historical monument or site (palaces, castles, churches, archaeological sites, gardens, etc.); HC3 – Visited a museum or gallery; HC4 – Attended a traditional event (e.g. food festival, carnival, puppet theatre, floral festival, etc.); HC5 – Visited a traditional craft workplace (e.g. weaving, glass blowing, decorative art, embroidery, making musical

instruments or pottery, etc.); HC6 – Been to the cinema or a film heritage festival to see a classic European film produced at least 10 years ago; and HC7 – Seen a traditional or classical performing arts event (e.g. music, including opera, dance or theatre, folk music, etc.).

3. Main barriers in experiencing cultural heritage (MB), with the following sub-variables: MB1 – Lack of interest; MB2 – Lack of time; MB3 – Cost; MB4 – Lack of information; MB5 – Lack or limited choice of cultural heritage sites or activities in the area; MB6 – Poor quality of cultural heritage sites or activities in the area; and MB7 – Cultural heritage sites or activities are too remote or difficult for you to access.

Associations between internet usage and frequency of attending cultural heritage-related activities, respectively internet usage and main barriers in experiencing cultural heritage have been measured using the Pearson correlation coefficient.

3. Main results

More than half (55 %) of the European consumers have used the internet for cultural heritage related purposes which may suggest a relatively sound presence of the network of networks in the daily life of the European cultural heritage consumer. At a more careful evaluation, there are two areas requiring significant improvements: the differences between the European countries, respectively the purposes of this usage. There is a large spread of the internet usage for cultural heritage purposes frequency: from 24 % (Portugal), 37 % (Greece) and 41 % (Bulgaria) to 83 % (Netherlands), 84 % (Belgium) or 85 % (Sweden). There are also significant differences in terms of the purposes: almost a third (31 %) of the European consumers use internet to prepare a visit or a holidays by looking up general information related to cultural heritage, about one in five use it to buy or book related services (23 %), view cultural heritage-related content experiencing or preparing the experience (21 %), know more about a cultural heritage site or event after experiencing (19 %), and around one in ten use it to create or share cultural heritage-related content (11 %) or give their opinion of a cultural heritage site or activity (6 %).

Still, the internet usage supports the consumption of cultural heritage (see Table 1): using internet for at least one cultural heritage related purpose associates significantly ($p < .001$) and positively with the specific forms of consumption represented by visiting a traditional craft workplace, a museum or gallery, a library or archive, a historical monument or site, and seeing a traditional or classical performing arts event. Positive, less intense and not statistically significant associations have been measured in the cases of going to a cinema or film heritage festival and attending traditional events. The internet proves to be supportive to the public of the traditional craft workplaces, museums and galleries, libraries and archives, historical monuments and sites, and traditional or classical performing arts events, and rather informative to the audiences of film heritage festivals and/or traditional events.

Table 1. Measures of associations between the usage of the internet and forms of cultural heritage consumption in the European Union.

	IU0	IU1	IU2	IU3	IU4	IU5	IU6
HC1	0.765***	0.800***	0.801***	0.720***	0.563**	0.738***	0.505**
	< .001	< .001	< .001	< .001	0.002	< .001	0.006
HC2	0.744***	0.875***	0.856***	0.812***	0.623***	0.779***	0.484**
	< .001	< .001	< .001	< .001	< .001	< .001	0.009
HC3	0.847***	0.908***	0.932***	0.864***	0.691***	0.819***	0.542**
	< .001	< .001	< .001	< .001	< .001	< .001	0.003
HC4	0.291	0.361	0.357	0.222	0.186	0.215	0.288
	0.133	0.059	0.062	0.256	0.345	0.271	0.137
HC5	0.820***	0.837***	0.867***	0.774***	0.633***	0.783***	0.589***
	< .001	< .001	< .001	< .001	< .001	< .001	< .001
HC6	0.524**	0.290	0.470*	0.310	0.252	0.317	0.283
	0.004	0.135	0.012	0.109	0.195	0.100	0.145
HC7	0.661***	0.706***	0.766***	0.639***	0.482**	0.635***	0.506**
	< .001	< .001	< .001	< .001	0.009	< .001	0.006

* p < .05, ** p < .01, *** p < .001

Internet usage also stimulates the frequency of experiencing different forms of cultural heritage. Using more intensely the internet for at least one cultural heritage related purpose associates significantly ($p < .001$) and positively with the frequency of visiting museums and galleries, historical monuments or sites, seeing traditional or classical performing arts events, and visiting libraries or archives. The internet is supportive but not influencing significantly the frequency of experiencing in the cases of visiting traditional craft workplaces, going to the cinema or a film heritage festivals or attending traditional events.

The digital marketing funnel works almost perfectly (a significant association of $p < .001$ for all six ways of internet usage) in the case of experiencing traditional craft workplaces. The public looks up for general information related to the accessibility, facilities and main features of these traditional craft workplaces and, particularly, to buy or book the related cultural heritage output – goods, services, brands, events or activities. The Internet is also used to improve the experience by knowing more about the traditional craft workplaces after and viewing related content during the visit. The Internet is least yet significantly used to

create or share content and give post-experience opinions about the related heritage content.

Visiting historical monuments or sites, as well as museums or galleries, have the same pattern of internet usage. Heritage consumers interested in palaces, castles, churches, archaeological sites, gardens, museums, and galleries employ the internet to look up general information related to the sites they are about to experience during their visits or holidays and/or to buy or book related cultural heritage output. They add value to their experiences by accessing and viewing cultural heritage-related content during their visits and improve them by getting more information about the sites after the visit. The Internet is less used to create or share cultural heritage-related content and to give opinions about heritage sites or activities.

Visiting libraries or archives and seeing traditional or classical performing arts events reveal a slightly different pattern of internet usage. Their public use the internet to looking up general information related to the cultural heritage content they would like to experience, buy or book specific cultural heritage output, access related content to during experiencing the heritage and knowing more about the heritage entity and/or content after experiencing it. The Internet is less used to create and share cultural heritage-related content or to give opinions about cultural heritage sites or activities.

The digital marketing funnel seems less relevant to the heritage consumers interested in going to cinema or film heritage festivals and, particularly, attending traditional events. Internet is used to look up information related to the such as cinema, film or food festivals, carnivals, puppet theatres, floral festival, buying or booking specific cultural heritage output, viewing cultural heritage-related content during experiencing or knowing more about this heritage after experiencing, creating or sharing later cultural heritage-related content or even giving opinions about cultural heritage sites or activity, but its employment does not associates significantly with the consumption of these forms of cultural heritage. The exception from the rule is represented by the buying or booking related cultural heritage output in the case of cinema or a film heritage festivals.

The barriers invoked by the nine out of ten heritage consumers explain the modest cultural heritage consumption in the European Union. According to the Special Eurobarometer 466 (2017), lack of time (indicated by 37 % of the European consumers), cost (34 %), and lack of interest (31 %) together with the lack of information (25 %) completes the harmful set of factors reducing significantly the willingness to discover, explore, experience and enjoy a somehow worthless, somewhat expensive, rather unknown and boring enough heritage. Lack or limited choice (12 %), remoteness (12 %), and poor quality (6 %) of the cultural heritage sites are peripheral but reinforcing reasons of the reserved attitude towards the cultural heritage related opportunities.

Can a digital marketing funnel soften these barriers and improve the consumption of cultural heritage? Measures of the associations between internet usage and main barriers of the

cultural heritage consumption (see Table 2) suggest a partially positive answer.

Table 2. Measures of associations between the usage of the internet and barriers of cultural heritage consumption in the European Union.

	IU0	IU1	IU2	IU3	IU4	IU5	IU6
MB1	-0.214	-0.384*	-0.249	-0.310	-0.205	-0.322	-0.283
	0.274	0.043	0.201	0.109	0.296	0.095	0.144
MB2	0.215	0.182	0.154	0.169	-0.030	0.285	-0.076
	0.271	0.354	0.435	0.390	0.880	0.142	0.700
MB3	-0.480**	-0.608***	-0.580**	-0.614***	-0.593***	-0.521**	-0.459*
	0.010	< .001	0.001	< .001	< .001	0.004	0.014
MB4	-0.105	-0.259	-0.110	-0.128	-0.115	0.013	0.043
	0.594	0.184	0.578	0.517	0.561	0.949	0.828
MB5	-0.341	-0.448*	-0.312	-0.363	-0.250	-0.301	-0.113
	0.076	0.017	0.106	0.058	0.200	0.119	0.566
MB6	-0.407*	-0.565**	-0.462*	-0.472*	-0.239	-0.450*	0.028
	0.032	0.002	0.013	0.011	0.221	0.016	0.889
MB7	-0.150	-0.141	-0.040	-0.106	-0.197	0.042	0.189
	0.447	0.475	0.839	0.592	0.316	0.832	0.336

* p < .05, ** p < .01, *** p < .001

Using the internet to obtain relevant, sufficient, and attractive general information about the cultural offer and the cultural heritage output could significantly reduce the barriers of costs, poor quality, lack or limited choice and lack of interest. Buying or booking cultural heritage output, viewing related content during experiencing as well as knowing more about the cultural heritage related content after experiencing it, may significantly diminish the barriers represented by the cost and poor quality of the heritage sites supporting the feeling that the experience was worth it and provided a good value for money. Last but not least, creating or sharing cultural heritage-related content and giving opinions about a cultural heritage site or activity may also diminish the barrier represented by the cost associated with experiencing the cultural heritage. Cost appears as a common denominator for all the purposes of using the internet in connection with cultural heritage.

4. Conclusions

Given the extent of the internet access and usage, costs of running online communication campaigns, and comfort of interacting online with the consumers, providing online information and support about the organization and its products, services, brands, events or activities is the most obvious, convenient, but also simple way to connect to its public. Cultural entities make no exception from this evidence and perhaps the most important conclusion of this study is that the internet "does well" to the consumption of cultural heritage outputs: more internet means more cultural heritage consumption and integrating a digital component into the marketing funnel is a must.

Considering the stages to be developed in order to support the market presence of the cultural heritage output (discovery, exploration, experience and enjoy), the digital component of the marketing funnel seems to impact significantly all mostly by limiting the barriers of cultural consumption, especially the price. In fact, not the price in itself, but the value for money (and time!) received and appreciated as such by the heritage consumer. Although it may sound too commercial, we need to understand that all the commodities providers (including of cultural heritage output), compete nowadays for the budget and time of each consumer and designing and employing a marketing funnel with a consistent digital component is the most effective solution for the capitalization of the cultural heritage.

References

Eigenraam, A. W., Eelen, J., van Lin, A., Verlegh, P. W. J. (2018). A Consumer-based Taxonomy of Digital Customer Engagement Practices. *Journal of Interactive Marketing*, 44, 102-121.

European Commission (2017). *Special Eurobarometer 466 - October 2017 Cultural Heritage Report*. Retrieved from http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/ResultDoc/download/DocumentKy/80882.

International Council on Monuments and Sites. ICOMOS International Cultural Tourism Committee (2002). *ICOMOS International Cultural Tourism Charter. Principles And Guidelines For Managing Tourism At Places Of Cultural And Heritage Significance*. Retrieved from http://www.icomos.no/wp-content/uploads/2014/04/ICTC-Charter.pdf.

Kannan, P. K., Li, H. A. (2017). Digital marketing: A framework, review and research agenda. *International Journal of Research in Marketing*, 34, 22-45.

Leeflang, P. S. H., Verhoef, P. C., Dahlström, P., Freundt, T., (2014). Challenges and solutions for marketing in a digital era. *European Management Journal*, 32, 1-12.

Pasikowska-Schnass, M. (2017). *Access to culture in the European Union*. Retrieved from https://pagina.jccm.es/europa/pdf/PUBLICACIONES/2017%20access%20to%20culture%20EU.pdf.

Sabatini, F. (2019). Culture as fourth pillar of sustainable development: Perspectives for integration, paradigms of action. *European Journal of Sustainable Development*, 8 (3), 31-40.

Thurley, S. (2005). Into the future. Our strategy for 2005-2010. *Conservation Bulletin*, 49, 26-27.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11632

Combining content analysis and neural networks to analyze discussion topics in online comments about organic food

Hannah Danner[1], Gerhard Hagerer[2], Florian Kasischke[2], Georg Groh[2]
[1]TUM School of Management, Technical University of Munich, Germany, [2]TUM Department of Informatics, Technical University of Munich, Germany.

Abstract

Consumers increasingly share their opinions about products in social media. However, the analysis of this user-generated content is limited either to small, in-depth qualitative analyses or to larger but often more superficial analyses based on word frequencies. Using the example of online comments about organic food, we investigate the relationship between qualitative analyses and latest deep neural networks in three steps. First, a qualitative content analysis defines a class system of opinions. Second, a pre-trained neural network, the Universal Sentence Encoder, analyzes semantic features for each class. Third, we show by manual inspection and descriptive statistics that these features match with the given class structure from our qualitative study. We conclude that semantic features from deep pre-trained neural networks have the potential to serve for the analysis of larger data sets, in our case on organic food. We exemplify a way to scale up sample size while maintaining the detail of class systems provided by qualitative content analyses. As the USE is pre-trained on many domains, it can be applied to different domains than organic food and support consumer and public opinion researchers as well as marketing practitioners in further uncovering the potential of insights from user-generated content.

Keywords: *deep neural networks; natural language processing; consumer research; content analysis; social media; organic food.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

Novel communication technologies sparked the desire of users to publicly share opinions on online platforms (Ziegele et al., 2014). These developments provide an increasing amount of user-generated content, such as online user comments, which can be exploited by marketing and consumer research to gain insights into consumer thinking (Balducci & Marinova, 2018). Beginning with Kozinets' (2002) netnography of online communities, social scientists have increasingly analyzed textual user-generated content with established methods such as content analysis (Krippendorff, 2019). However, due to time and human resources required, such qualitative analyses are limited to small data samples. More recently, advances in automated text analysis and data collection enable consumer researchers to efficiently analyze larger datasets in a short amount of time and facilitate the detection of patterns, and compare measurements over time or between datasets. For an overview of methods see Berger et al., 2020). Frequently employed methods are dictionary-based approaches (e.g., LIWC, Tausczik & Pennebaker, 2010) relying on word frequencies. Researchers using automated text analysis have started to incorporate methods from the field of natural language processing (NLP, such as of data-mining, data-preprocessing, simple classifiers, and topic models (Latent Dirichlet Allocation, Blei, 2012) (for an overview see Vidal et al., 2018). However, to the best our knowledge, there has been little research on how qualitative and NLP methods can be combined fruitfully. Latest advances in NLP are neural networks that account for the semantic context of words, i.e., word embeddings (Mikolov et al., 2013), or sentences, i.e., sentence embeddings (Cer et al., 2018). In this paper, we explore how such embeddings particularly lend themselves to be combined with qualitative text analysis by matching the analysis-depth of the latter with the scope of pre-trained sentence embeddings. In three steps, we present a novel approach for how a qualitative content analysis can be combined and enhanced with deep neural networks for semantic similarity.

We apply the approach to the case of organic food. Not only is a growing share of consumers aware of and buys organic food (Hemmerling et al., 2015)—making it an increasingly important consumer research topic—, consumers also voice their opinions about organic food online (Danner & Menapace, 2020; Meza & Park, 2016; Olson, 2017). The analysis of online user-generated content can thus deliver valuable insights into which product attributes and related topics matter to consumers and what could be potential purchase drivers and barriers.

2. Methodology

In step 1 of our approach, a qualitative text analysis is conducted to develop a class system and manually classify a dataset of interest. In Step 2, we use semantic features from pre-trained neural networks to investigate the semantic characteristics and the respective frequencies for each class. Step 3 presents criteria to combine results of both methods.

2.1. Step 1 – Qualitative Analysis

To exemplify the approach, for step 1, we draw on a recent qualitative content analysis by Danner and Menapace (2020) of online comments about organic food. They manually extracted and classified consumer opinions (referred to as beliefs) about organic food to understand consumers' perception of organic. The authors collected 1069 online comments about organic food from high-coverage US news websites (e.g., nytimes.com, washingtonpost.com) and forums (e.g., reddit.com, quora.com). The 1069 comments consisted of 5510 sentences. Among these 5510 sentences, the two coders identified 1065 containing belief statements about organic food and subsequently classified those belief statements into 64 belief classes and 21 superordinate topics. For example, the sentence stated by a commenter *organic farming is better for nature* was attributed to the belief class *organic farming protects the environment*, which in turn was attributed to the topic class *environment*. By counting the frequencies of belief statements per category, the authors presented a detailed picture of topics salient to the online commenters in the data.

2.2. Step 2 – Universal Sentence Encoder

Using the same data and class system as in step 1, we find similar sentences for each class using the Universal Sentence Encoder (USE). USE is a recent advance in NLP and deep learning (Cer et al., 2018). Its architecture is based on the widely adopted Transformer architecture (Vaswani et al., 2017). USE is a deep neural network model pre-trained on large scale text corpora from many domains. From there, the statistical knowledge in terms of generalizable, intermediate, semantic vector representations, which are also referred to as features or embeddings, can be used to quantify the semantics of specific domains, here organic food. USE works on sentence level providing sentence embeddings. The semantics of a given sentence are expressed by its vector representation. When compared to other sentences, the cosine similarity ranges between 1 (similar) to -1 (dissimilar).

We applied USE to automatically find semantically similar sentences for each of the 64 beliefs identified by Danner and Menapace (2020) (e.g., *organic farming protects the environment*) (Table 1). First, USE transformed each of the 64 beliefs and the 5510 sentences into an embedding. Second, USE measured the cosine similarity, i.e. the angular distance, between the embedding of each of the 64 beliefs (also referred to as seed sentences) and each of the 5510 sentences. When choosing a low threshold level for cosine similarity (i.e., the closer to -1), many sentences are considered as similar, whereas at high levels fewer sentences are considered as similar.

2.3. Step 3 - Evaluation

Eventually, we determine the appropriate level of semantic similarity, i.e., the respective cosine similarity threshold level which yields similar frequencies compared to the qualitative

content analysis as reference. To this end, we inspect the thresholding results for cosine similarity levels from 0.7 to 0.84 based on the following criteria. (1) In the content analysis, 1065 sentences were relevant as in containing beliefs about organic food. A meaningful sentence filtering should yield a similar amount of relevant sentences. (2) The number of sentences assigned into the different classes should be similar for both methods. Therefore, we inspected the relative class frequencies and also calculated the Pearson correlation between the class frequencies for different cosine similarity levels. Figure 1 displays a trade-off between semantic similarity and class frequencies: the lower the cosine similarity (i.e., the less similar the sentences), the higher the correlation between the two methods. (3) Manual inspection should confirm the semantic cohesion between the manually and the automatically assigned sentences. Note that we performed the evaluation at topic level (21 topic classes) as the 64 belief classes are very detailed and in part semantically too similar (e.g., *organic farming is better for the environment* and *conventional farming harms the environment*).

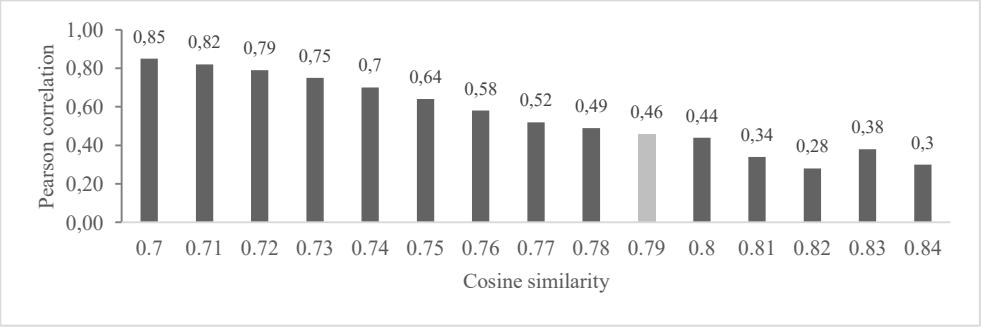


Figure 1. Pearson correlation of class frequencies (21 topic classes) between content analysis and USE. Source: own illustration.

3. Results

Applying the aforementioned evaluation criteria, the thresholding performed best at a cosine similarity of 0.79. (1) At this level of similarity, USE found 1376 relevant sentences, which roughly corresponds to the 1065 relevant sentences identified in the manual analysis. (2) As highlighted in Figure 1, for cosine similarity of 0.79, both methods yielded similar class frequencies, indicated by a correlation of r = 0.46. However, class frequencies do not match perfectly. Looking at the relative class frequencies for each of the 21 topic classes in Figure 2, we find that the class frequencies for both methods are more similar for some topics than for others. For example, the topic *environment* accounts for 11% of sentences in the content analysis and 18% in the similarity thresholding. The most frequent topics in the content analysis were *system integrity*, *food safety*, *environment*; the most frequent topics in USE were *environment*, *system integrity*, *farmer welfare*.

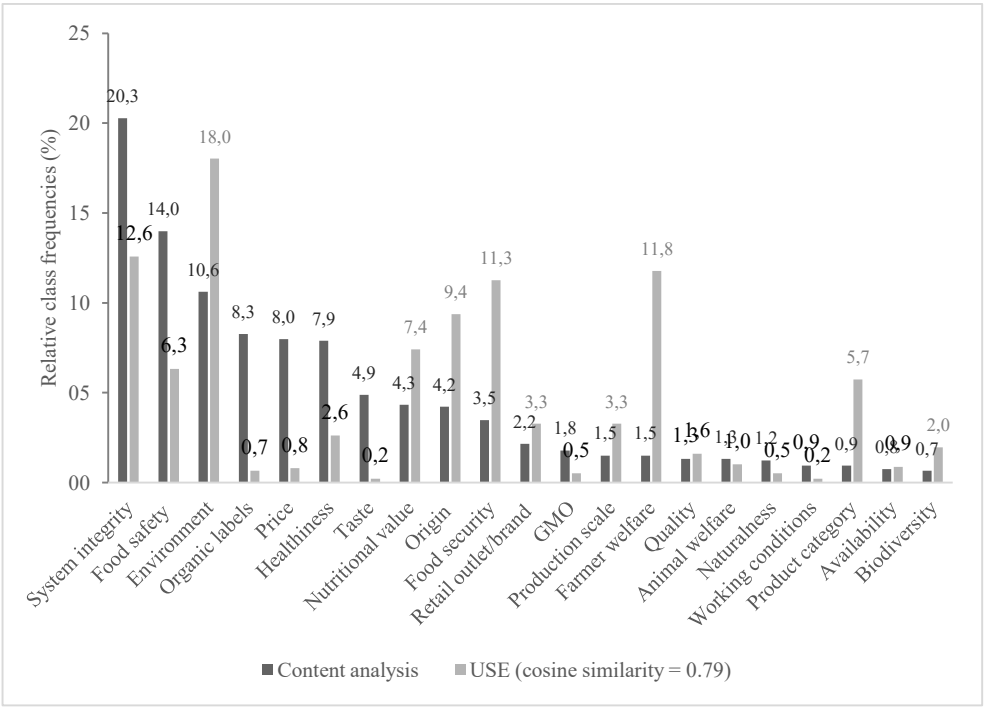


Figure 2. Relative class frequencies (21 topic classes) in content analysis and USE.
Topics are ordered in descending frequency according to the content analysis. Source: own illustration.

Table 1. A seed sentence from content analysis and the 11 sentences identified as similar by USE (cosine similarity = 0.79).

seed	Organic farming protects the environment.
1	Organic farming can help to preserve our environment for future generations.
2	The depletion of the soil and monoculture is what causes factory farming produce to be less nutritious than organic.
3	Mythbusting 101: Organic Farming > Conventional Agriculture
4	A lot of what I've read has said that organic farming is not better for the environment.
5	Organic is for the environment.
6	And from this we hear that organic farming is "devastating" to the environment.
7	Organic farming is much closer to the way Mother Nature farms.
8	GMOs can be super beneficial - to the consumer, the farmer, the environment.
9	Organic farming is greener
10	Besides delivering health benefits, organic farming is better for the environment.
11	Organic is for the environment.

Source: own illustration.

(3) For cosine similarity of 0.79, manual inspection showed very high semantic cohesion between the seed sentences per topic and the sentences identified as similar by USE. Table 1 displays the 11 sentences that USE found to be similar to the belief *organic farming protects the environment* at a cosine similarity of 0.79. All 11 are concerned with the effect of organic farming on the environment. However, sentences 3, 4, and 6 carry negative and thus the sentiment opposite to the seed sentence. Thus, while USE correctly identifies the topic, the sentiment is not always correctly classified, which is one reason why comparisons at topic level were chosen for this study. In addition, the manual inspection of the sentences classified by both methods proved that both methods classified largely the same sentences in the respective classes.

4. Discussion

USE appears to be an effective and easy to use method to analyze large text corpora by searching for sentences that are semantically similar to seed sentences of interest. Seed sentences can originate, for instance, from a small-scale qualitative study—here the belief classes identified by Danner & Menapace (2020). Provided a manually developed class system, it can analyze any unseen dataset, —here 5510 sentences on organic food—,

according to semantic similarity. In the present example, a human researcher selected the required level of similarity by evaluating the features generated by USE based of descriptive statistics and manual inspection. We suggested several criteria to select the appropriate similarity level as an alternative to training a classifier. Training a reliable classifier to classify fine-grained classes as complex as 64 different organic food beliefs requires large amounts of labeled data, which often exceed the resources of common research projects in the field of consumer and opinion research, and as it also applied to the presented example.

The selected similarity threshold was valid as the filtered sentences were widely coherent with the qualitative content analysis. In a subsequent step, USE could be applied to filter a larger unseen data set on organic food. Thus, the potential of the suggested approach lies in its scalability. We can extrapolate the detail of insight characteristic of qualitative research to analyze class frequencies in a larger data set of user-generated content.

Being still in an early phase, our approach bears potential for further refinements. We used a very large class system with 64 belief classes grouped into 21 topics, which also contained classes semantically very similar to each other. Using fewer and more distinct classes could thus improve the coherence between a manual classification and automatic classification based on USE. Furthermore, USE reliably finds the sentences containing similar topics, but does not always correctly distinguish positive and negative sentiment regarding the topic. Therefore, while suitable for topic classification, its use for sentiment analysis is bound to the manual control of a human researcher and domain expert. The imperfect match between manual classification and automatic filtering may also originate from the selection of the unit of analysis, a well-discussed issue in qualitative research (Campbell et al., 2013). The unit of analysis in USE are sentences, whereas in the content analysis, the unit of analysis could also stretch beyond a single sentence, and qualitative researchers can use domain knowledge for understanding and classifying text.

5. Conclusion

In a three-step approach, we suggested how a topic classification of a qualitative content analysis—here of online comments about organic food—can be combined with neural networks like USE to find similar sentences. We proved that embedding techniques largely fit the results of qualitative analysis and point out their methodological potential. USE considers the semantic coherence between words and sentences and delivers in-depth insights by providing the original consumer phrasings (see Table 1) instead of abstract word lists and word frequencies as in more simple approaches of automated text analysis, such as dictionary-based approaches or LDA topic modeling.

Additional potential lies in cross-lingual applications using multilingual USE: Researchers can use the same seed sentences in one language and analyze data sets in different languages

to make cross-country comparisons. Analyzing user-generated content, consumer researchers can learn about which product attributes and topics salient to consumers and potentially serve as purchase drivers or barriers. Based on this, consumer typologies and clusters can be derived. An improved understanding of consumers' opinions can support the design of organic products as well as labeling policies. Another application of USE lies in using items of established scales from survey research as seed sentences and analyze their similarity and prevalence in social media data. In addition, the suggested approach could be promising for market monitoring based on the targeted detection of social media content. For example, social media managers can observe the prevalence and development of certain opinions over time.

References

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557–590. https://doi.org/10.1007/s11747-018-0581-x

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing*, *84*(1), 1–25. https://doi.org/10.1177/0022242919873106

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding In-depth Semistructured Interviews. *Sociological Methods & Research*, *42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., . . . Kurzweil, R. (2018). Universal Sentence Encoder. *ArXiv*. Retrieved from http://arxiv.org/pdf/1803.11175v2

Danner, H., & Menapace, L. (2020). Using Online Comments to Explore Consumer Beliefs Regarding Organic Food in German-Speaking Countries and the United States. *Food Quality and Preference*, *83*(103912). https://doi.org/10.1016/j.foodqual.2020.103912

Hemmerling, S., Hamm, U., & Spiller, A. (2015). Consumption behaviour regarding organic food from a marketing perspective—a literature review. *Organic Agriculture*, *5*(4), 277–313. https://doi.org/10.1007/s13165-015-0109-3

Kozinets, R. V. (2002). The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities. *Journal of Marketing Research*, *39*(1), 61–72. https://doi.org/10.1509/jmkr.39.1.61.18935

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (Fourth edition). Los Angeles, London, New Delhi, Singapore: SAGE.

Meza, X. V., & Park, H. W. (2016). Organic products in Mexico and South Korea on Twitter. *Journal of Business Ethics*, *135*(3), 587–603. https://doi.org/10.1007/s10551-014-2345-y

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Retrieved from http://arxiv.org/pdf/1301.3781v3

Olson, E. L. (2017). The rationalization and persistence of organic food beliefs in the face of contrary evidence. *Journal of Cleaner Production*, *140*, 1007–1013. https://doi.org/10.1016/j.jclepro.2016.06.005

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. *ArXiv*. Retrieved from http://arxiv.org/pdf/1706.03762v5

Vidal, L., Ares, G., & Jaeger, S. R. (2018). Chapter 6 - Application of Social Media for Consumer Research. In G. Ares & P. Varela (Eds.), *Woodhead Publishing Series in Food Science, Technology and Nutrition. Methods in consumer research* (pp. 125–155). Duxford, United Kingdom: Woodhead Publishing.

Ziegele, M., Breiner, T., & Quiring, O. (2014). What Creates Interactivity in Online News Discussions? An Exploratory Analysis of Discussion Factors in User Comments on News Items. *Journal of Communication*, *64*(6), 1111–1138. https://doi.org/10.1111/jcom.12123

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11633

Setting Crunchbase for Data Science: Preprocessing, Data Integration and Feature Engineering

Francesco Ferrati, Moreno Muffatto
School of Entrepreneurship (SCENT), Department of Industrial Engineering, University of Padova, Italy.

Abstract

In order to support equity investors in their decision-making process, researchers are exploring the potential of machine learning algorithms to predict the financial success of startup ventures. In this context, a key role is played by the significance of the data used, which should reflect most of the variables considered by investors in their screening and evaluation activity. This paper provides a detailed description of the data management process that can be followed to obtain such a dataset. Using Crunchbase as the main data source, other databases have been integrated to enrich the information content and support the feature engineering process. Specifically, the following sources has been considered: USPTO PatentsView, Kauffman Indicators of Entrepreneurship, Academic Ranking of World Universities, CB Insights ranking of top-investors. The final dataset contains the profiles of 138,637 US-based ventures founded between 2000 and 2019. For each company the elements assessed by equity investors have been analyzed. Among others, the following specific areas were considered for each company: location, industry, founding team, intellectual property and funding round history. Data related to each area have been formalized in a series of features ready to be used in a machine learning context.

Keywords: *Crunchbase; startup; investments; feature engineering; data mining; machine learning.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

The large amount of business-related data available today, allows researchers in entrepreneurship, economics and social sciences to investigate complex phenomena using innovative approaches. In an economic system where entrepreneurship activities are considered as a driver for growth and social improvement, a very topical issue concerns the possibility of predicting to some extent the probability of success of early stage technology-driven ventures. Due to their inherently high level of innovation, startup companies are considered to be highly uncertain and risky business activities and the statistics on their failure rate are still very high (Gage, 2012). From an academic point of view, since the 1980s researchers have been analyzing the equity investors' decision-making process, questioning about its effectiveness and wondering whether it could be improved (MacMillan, Siegel & Narasimha, 1985). The assumption underlying this research stream is that the use of more effective assessment criteria could lead to the identification of the best entrepreneurial projects, which might in turn contribute to the success of an investment portfolio (Zacharakis & Meyer, 1998).

Leveraging the growing amount of available data as well as the increasing accessibility of advanced data mining frameworks, in recent years researchers have started exploring new approaches to the so called "picking winners" problem. Specifically, they have begun investigating the potential of machine learning algorithms as a tool to support venture capitalist in their screening and evaluation processes. Given the complexity of the task, retrieving and processing data that can be used to properly model an early stage venture has a huge impact on the performance of the final models. In this regard, Crunchbase is an innovative online platform collecting and providing business information about technology-driven companies, investors, funding rounds and key people involved in the entrepreneurial network (Ferrati & Muffatto, 2020). Thanks to the quantity and quality of their data, it is effectively used not only by practitioners (e.g., entrepreneurs, investors or policy makers), but also by academic researchers who intend to apply quantitative approaches to the research on entrepreneurship and innovation. In this context, a key problem that can be addressed by applying machine learning algorithms to Crunchbase data concerns the prediction of a company's exit event, commonly considered as the critical milestone defining a company's financial success. (Krishna, Agrawal & Choudhary, 2016).

Although previous works on this topic have usually described the data modeling process in detail, pre-processing, data integration and feature engineering activities have not always been covered in depth. Specifically, the logic used for the identification and selection of the features used in the models as well as the steps followed to obtain the considered samples have generally not been fully described. As a result, the considered datasets have not always been clear in their content and the models' results could be therefore hard to interpret. In addition, in previous contributions Crunchbase has generally been used as the only data

source, and no activity has been carried out to integrate other databases to enrich the information content. The present work aims to fill these two research gaps by providing a full example of how the database can be prepared according to the well established-steps of the data science workflow.

This paper is organized as follow. Section 2.1 defines the research goal for which the database can be used. Section 2.2 reports the relevant data sources that have been integrated into Crunchbase in order to add some key features. Section 2.3 describes in full detail all the steps that have been followed to create the final dataset. Finally, Section 3 discuss the obtained results and presents some elements for future research.

2. Steps of the data setting process

The setting of the dataset was carried out following the main steps of the data science process. After carrying out a literature review of the assessment criteria used by equity investors (Ferrati & Muffatto, 2019), and identifying the key information to evaluate a startup company, we defined the purpose for which the database can be used. Considering Crunchbase as the main data source, we then search for the most useful publicly available data to enrich its information content. We then moved on to the data preparation phase, going to combine, clean and transform the available data. Once aggregated the different datasets, we then made an analysis of the available variables, performing a feature selection and feature engineering activity. Figure 1 shows in bold the steps presented here in the context of a data science workflow.

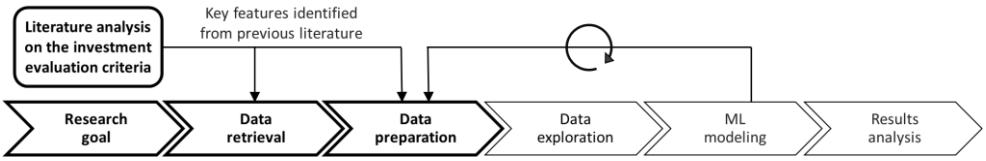


Figure 1. The six steps of the data science workflow.

2.1. Research goal definition

In order to prepare data effectively, the first step is to clearly define the research goal for which the dataset will be used. The dataset we prepared can be used in a supervised machine learning environment to predict the financial success of a startup company. In this context, a venture's financial success is defined by the occurrence of an exit event, both in the form of merger and acquisition (M&A) and an Initial Public Offering (IPO). Crunchbase provides information about the status of a company through a specific categorical variable that can assume four different values: operating, closed, acquired or IPO. The “status” can be effectively used as a target variable in a machine learning classification problem.

2.2. Data retrieval

Since a machine learning model should support investors in their screening and evaluation process, the largest number of significant variables should be identified in the data retrieval phase to cover most of the aspects generally considered by investors. Crunchbase itself provides many useful information. The database is organized in seventeen .csv files: *Organizations, Organization descriptions, Category groups, Funding rounds, Investors, Investments, Investment partners, Funds, People, People descriptions, Jobs, Degrees, Acquisitions, Ipos, Organization parents, Event appearances* and finally *Events.* By grouping the information contained in each individual dataset, the complete database covers five macro information areas respectively related to organizations, investment activities, people, exits and public events. In order to map the content and give a representation of how the different datasets are linked together, we started by inferring the Crunchbase relationships scheme as shown in Figure 2. After an accurate exploration of the individual files, we select the datasets (colored in black) that could provide the most relevant information according to the considered research goal. Despite the large amount of information already provided by Crunchbase, we decided to go further and integrate it with other additional data sources in order to enrich the information content. Specifically, as shown in Figure 2, four additional data sources have been considered.

- **United States Patents**. Because of the high level of innovation experienced by technology-driven startups, a key competitive advantage concerns their intellectual property portfolio. Since Crunchbase does not directly provide this kind of information, we used the public data collected by the USPTO PatentsView platform to search for patents assigned to each company. A similar process has been reported also in previous literature considering the PATSTAT dataset (Menon & Tarasconi, 2017).
- **Kauffman Indicators of Entrepreneurship**. For each company in the database, Crunchbase provides its location in terms of country, state, region and city. Since the entrepreneurial ecosystem has a strong impact on a company's performance (Sheriff & Muffatto, 2018), the values of the Kauffman Indicators of Entrepreneurship have been integrated, providing some key metrics for each US state.
- **Academic Ranking of World Universities**: In order to assess the educational background of each company's team members, we integrated the ARWU 2018 ranking to identify the founders who graduated from a worldwide top university.
- **CB Insights top investors list**. The fact that a successful investor decides to support a startup has a strong impact in terms of credibility and can facilitate the occurrence of subsequent funding rounds. In order to identify the companies backed by top investors, we merged the investments' data with the ranking made by CB Insights

in 2019. This list provides the 48 top-investors who have backed the most unicorn companies.

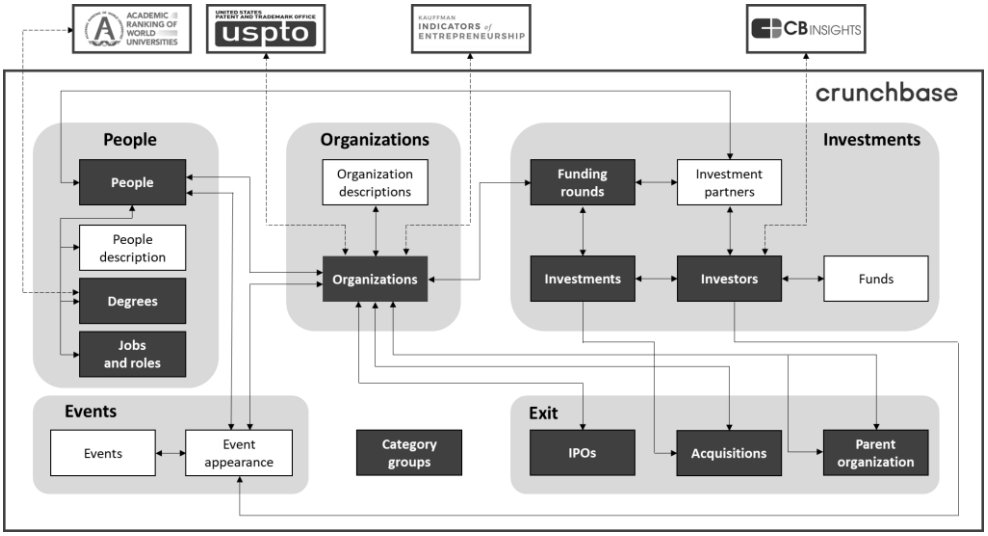


Figure 2. Scheme of the relationships between the Crunchbase datasets and the additional integrated sources.

2.3. Data preparation

The version of Crunchbase considered in this work is dated May 21, 2019 and contains information about 760,590 organizations, 121,509 investors (e.g., venture capital, angel investors, etc.), 263,426 funding rounds, 890,429 people, 1,346,357 jobs, 17,068 IPOs and 89,959 acquisitions. Starting from the raw data, the final dataset consists in a single file and considers the company as unit of analysis. Each row corresponds to a different company and more than 130 columns provide the features that can be used for machine learning. All the steps that led to the creation of the final dataset are described below in sequential order.

1. **Add exit info to organizations**. Starting from the "organizations" dataset, the information about exit events was integrated from the "acquisitions" and "ipos" datasets. Whereas an organization was acquired several times or reported more than one IPO, only the first exit event was considered. Among the 760,590 organizations, 85,099 have been acquired and 16,457 went public. A boolean value was used to label the organizations that made an exit and the target variable was so defined.
2. **Add lifetime info to organizations**. In order to take into account the phase in the business life cycle of each company, a variable has been added for the computation of the company's lifespan (in number of months). For the organizations with status equal to "acquired", "ipo" or "closed", the difference between the date of these events and the foundation date was considered. For the still "operating" companies,

the difference between 2018-12-31 and the date of foundation was considered. Organizations with missing dates were excluded. After this step, 518,661 organizations remained.

3. **Filtering companies**. Since Crunchbase considers as organizations both companies, investment firms and universities/schools, the dataset was filtered considering only companies (489,710 remaining companies). To focus on a specific entrepreneurial ecosystem, only US-based companies were then considered (195,542 remaining companies) and in order to not suffer the effects of "dot-com bubble", only companies founded between 2000 and 2019 have been finally considered (138,637 remaining companies).
4. **Founding team analysis**. Since the quality of the founding team is a key element for the success of a startup, for each company information about its founders has been integrated. In order to identify the people related to each company, the "jobs" and "people" datasets were used, selecting only the jobs related to the considered companies. Of the 1,346,357 records in the "jobs" dataset, 458,429 were related to the considered companies. On the other hand, 45,153 companies didn't report any job information, resulting in a lack of data about their founding team. In order to identify the founders of each company, only "founder" or "co-founder" job types were considered (84,834 unique founders were identified, some of whom had founded more than one company). In order to get a measure of the founders' work experience, all the jobs carried out in their career have been identified (195,993 jobs carried out by the founders have been identified). To understand the roles covered by founders, for each company the presence of some key chief-roles (e.g. CEO, CTO, COO, CFO, etc.) has been verified. As the "degrees" dataset provides information about the education of the registered persons, for each founder an analysis of the educational background was carried out. Of all the 335,414 degrees collected in the dataset, only those related to founders were extracted, resulting in 100,366 degrees. Each degree was classified according both to the type (e.g., bachelor degrees, master degrees, doctoral degrees, etc.) and to fifteen subject areas in order to associate to each company the team's areas of knowledge (e.g. business, engineering, computer science, science, etc.). For each degree, it was also verified whether the title was obtained from a university in the top 25 of the ARWU 2018 ranking. In total 11,783 degrees were obtained from a top university. Finally, for each company the gender of the founders has been considered (73,853 males, 10,586 females, 7 other and 46 not declared). In summary, the team analysis allowed to define for each company the following groups of features: number of founders (N.F.), N.F. per chief role, N.F. per type of degree, N.F. per subject degree, N.F. from top universities, N.F. per gender, average number of companies founded by each founder (serial entrepreneurs).

5. **Location analysis**. In order to enrich the information about the entrepreneurial ecosystem in which each company operates, we integrated Crunchbase with the five Kauffman indicators of entrepreneurship (i.e., Kauffman early-stage entrepreneurship index, rate of new entrepreneurs, opportunity share of new entrepreneurs, startup early job creation, startup early survival rate). Since each index assumes different values depending on the year and the US state, the association was made considering the state in which the company is located and the average (and median) value of the indicators calculated over the years of the company's lifespan.
6. **Sector analysis**. Crunchbase provides two variables, called "categories" and "category groups", for the classification of companies' activities. The variable "categories" can take one or more labels (related to industries, technologies, business models, etc,) from 680 possible options. To make these values more consistent with each other, we introduced a new classification scheme in order to reduce the 680 categories to 64 main areas. Then, we reclassified each company according to the 64 areas so identified.
7. **Investments analysis**. One of the most important information provided by Crunchbase regards the funding rounds collected by each company. The "funding rounds" dataset reports 263,426 rounds and the "investments" dataset collects 400,432 investments (the term *investment* refers to the participation of a single investor in a specific funding round). The information has been filtered considering only the rounds raised by the selected companies. As a result, we identified 52,037 companies with at least one funding round, 113,572 rounds related to them and 30,975 unique investors involved. For each company the number of rounds, the total amount collected (in USD) and the number of total as well as unique and serial investors were calculated. Finally, considering the top-investors ranking by CB Insights, for each company the number of top investors in portfolio was computed.
8. **Acquisitions analysis**. The Crunchbase "acquisitions" dataset collects information about 89,959 acquisitions. Starting from this data, the number of acquisitions made by the selected companies have been derived. As a result, we identified 12,223 acquisitions made by 6,369 companies among the considered one.
9. **Patent analysis**. Since the companies in the sample are all based in the USA, their patent portfolio was analyzed using the data collected by the USPTO PatentsView platform. Since both Crunchbase and PatentsView provide information about the location of each company and assignee, the match between the two datasets was made using a concatenate string between the company name and the U.S. state where it is based. In this phase special attention was paid to make the companies names homogeneous between the two datasets (e.g. by removing all the legal entity types abbreviations) and to manage homonymy cases. All the patents dated after a

company acquisition or IPO were excluded from the computation. The merge process identified 9,025 companies in the sample with at least one patent. For each of the identified companies, all the patents registered in the "patent" dataset were searched and 69,537 patents were identified. Patent data has been filtered considering only utility and design patents and patent applications were dropped from the computation. For each company the number of patents (utility and design) was reported.

Table 1 summarizes the content of the final dataset.

Table 1. Content of the final dataset.

	Number
Companies (based in USA and founded between 2000 and 2019)	138,637
Founders	84,834
Funding rounds	113,572
Companies with at least one funding round	52,037
Investors involved in the considered funding rounds	30,975
Companies having made at least one acquisition	6,369
Acquisitions made by the companies	12,223
Patents granted to the companies	69,537
Companies with at least one patent	9,025

3. Conclusion and future research

In this paper we presented the steps that could be followed to prepare Crunchbase to be used in machine learning to predict a startup's exit event. A series of filters and operations were applied to make the dataset as consistent as possible, and to integrate other information not considered in previous contributions. For future research, other data sources could be integrated to cover aspects related to products or services, business models, competitors and financials. Feature importance could be analyzed by applying logistic regression algorithm.

Acknowledgment

This research was made possible thanks to the support of Crunchbase Inc. http://www.crunchbase.com

References

Ferrati, F., & Muffatto, M. (2019). A Systematic Literature Review of the Assessment Criteria Applied by Equity Investors. *In 14th European Conference on Innovation and Entrepreneurship* 304-312

Ferrati, F., & Muffatto, M. (2020). Using Crunchbase for research in Entrepreneurship: data content and structure. *In 19th European Conference on Research Methodology for Business and Management Studies*

Gage, D. (2012). The venture capital secret: 3 out of 4 start-ups fail [online]. *The Wall Street Journal U.S. Edition*, updated Sept. 20, 2012 12:01 am ET.

Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the outcome of startups: less failure, more success. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* 798-805.

MacMillan, I. C., Siegel, R., & Narasimha, P. S. (1985). Criteria used by venture capitalists to evaluate new venture proposals. *Journal of Business Venturing*, 1(1), 119-128.

Menon, C., & Tarasconi, G. (2017). Matching Crunchbase with patent data. *OECD Science, Technology and Industry Working Papers*, 2017(7), 1-21.

Sheriff, M., & Muffatto, M. (2018). High-tech entrepreneurial ecosystems: using a complex adaptive systems framework. *International Journal of Entrepreneurship and Innovation Management*, 22(6), 615-634.

Zacharakis, A. L., & Meyer, G. D. (1998). A lack of insight: do venture capitalists really understand their own decision process? *Journal of Business Venturing*, 13(1), 57-76.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11635

Information balance between newspapers and social networks

Francesco Mazzeo Rinaldi[1], Andrea Russo[2], Giovanni Giuffrida[1]
[1]Department of Political and Social Science, University of Catania, Italy, [2]Department of Physics and Astronomy, University of Catania, Italy.

Abstract

Competing newspapers, tend to publish the same information in a given time frame. However, each editor tends to aggregate and present the news according to certain criteria such as editorial policies, filtering strategies, readers base, etc. Thus, the proper choice and filtering of information makes one newspaper different from the other and, the proper management of such criteria, may deem the success or failure of a newspaper.

From the editor's perspective, the news selection process is a trade-off between informativeness and attractiveness, as determined by the readership. Moreover, is it possible that cultural and political inputs from social media may impact the news selection process?

Political news on social networks represent nowadays a valuable informative asset that gives the possibility to correlate newspaper information with public request expressed on social networks. We believe that it is possible to develop a theory to mitigate the newspaper's cultural identity with the public information needs collected on social media.

In our work, we show how to measure the society's request for information through the analysis of public reaction to certain articles on social networks, in particular we present how studying the hashtags and articles shared can be conveyed to understand social dynamics in nowadays discussion.

Keywords: *Information; Newspaper; Social media, Social need; Big Data.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

The common goal of all newspapers is to make money by selling news. Thus, their news, and their presentation, must be attractive for the readership. However, considering the amount of new information constantly produced, a proper continuous smart selection of those is necessary and represent a crucial factor for the newspaper survival.

In this context, information extracted in real time from social media provides a valid input for the selection process. From a research perspective, data becomes crucial to understand how newspapers gain information to balance their cultural basis with the population's informative needs.

The purpose of this paper is to demonstrate that social network data could be successfully used as a new resource for newspapers. Many newspapers use big data to improve their articles, but in this case, we propose a new way to analyze big data to improve topic selection. It would be feasible that, with a more peculiar use of big data there will be the possibility to respond to unconscious requests for information (on a specific topic) by the social network's users. This unconscious need could be recognized and used in order to take advantage of that, in order to create highly informative content, so as to achieve the end of people's interest in the topic, even though the newspaper is specialized in a different sector.

From the methodologic point of view, hashtags can be used to estimate if there is apprehension or curiosity about a particular topic. For instance, we could study whether the word "Bojo", an acronym for Boris Johnson communicating closeness to the persons, is used more frequently among readers talking positively, about Brexit, compared to the ones not approving the Brexit plan. Thanks to this logic it is plausible to relate the use of the name to the newspaper's position on Brexit.

Today, every newspaper has its public digital space where readers discuss and share information. Thus, it becomes possible to enhance such a network by improving it or to attract new clusters of people that have a similar culture to the newspapers' one.

In particular, we can understand which topics are more suitable for the newspaper as recognized by the public. For example, Justice and Economics news are the most prominent themes for newspapers analyzing the political system, as well as wars and scientific discoveries.

By measuring real-time hashtag trends on social networks for particular topics we could promptly inform the newspaper about groundbreaking news.

For instance, if a newspaper selects the news about Brexit highlighting the economic effects rather than the causes of Brexit itself, assuming that to be the informational need of people if favor of Brexit, there is the possibility that the reader's request appears on the social networks.

It would be possible to analyze the selection process of the newspaper by studying the hashtags inside the articles published, in relation to the sharing network generated by the user activity.

Thus, social media data analysis becomes a key factor to understand readers' news needs to prepare a more attractive and accurate article set on a newspaper.

A case study for this subject could be the commentaries' data mining on articles that discuss a new law proposal of the Italian government discussed in late 2019. The analysis shows whether readers commenting the article are in favour or against it. (Francesco Mazzeo Rinaldi, Giovanni Giuffrida, and Tom Negrete).

2. Scientific results & methodology

Big Data and algorithms are nowadays a very common topic in social and political sciences. The initial debate about technology and social sciences have produced many publications with many interesting results. In many cases technology has produced some relevant suggestions to advance scientific research in those disciplines. Furthermore, through its continuous improvement in recent years, technology has helped to better understand our own society. *Mutatis mutandis.*

Thanks to social network data and its related algorithms, today there is the possibility to measure the unbalance between newspapers and social needs for information. There are some relatively new researches that using big data have brought innovative results. The "traditional" sentiment-analysis on a topic or an article, can be used by a newspaper to intensify the research on data-driven strategies to be used commercially. The benefits of sentiment-analysis have long been recognized, but with this research we want to show how big data, if used for a precise topic, could answer the need for information of the social network users, widening the research domain and its methodological settlement. A popular topic on social network could create a contrast between different points of view or show disinformation. In both cases, but more specifically when there is little or no information, through the use of big data it is plausible to detect the peoples' need for information (among supply and demand). The existence of the Points of consideration (which will be later examined in depth) is valuable to highlight the scarcity of information (or high number of discussion on a topic) on social media, and the absence of articles (supply) from newspapers. An accurate data analysis could show a demand for information, which will later be acquired from the newspapers and as a result, suitable articles be produced. For instance, it has been used to study trend topics in social networks, and the corresponding offer of information, thus showing whether there is an offer from newspapers.

Such a correlation provides the opportunity to estimate the public reaction to certain articles. For instance, in the political context, the algorithm can analyse keywords and hashtag used as a thematic aggregator, so to measure the level of attractiveness of a specific news.

2.1. Data source

We propose to use Twitter as the data source because of three crucial aspects: first of all, Twitter makes it easier to acquire data more than the other competitors. The text format is easily analysed because of its shortness (compared to other social networks such as Facebook or TikTok): you have just 280 characters to express your idea, so the data mining process is simpler. Furthermore, Twitter user base is very interesting: there are public figures and various experts, and all their discussions are represented by an hashtag.

Linking user ids on social media and on newspapers, in some cases, may happen because the id or the email is the same on both platforms. This could be a very useful piece of information for our analysis. It is possible that a newspaper can ask the id users on social media of their subscribers to understand what they think about some kind of hashtag or a particular information.

Moreover, users need to give permission to link their social media account with their newspaper one. GDPR legislation allow this typology of data gathering, and the data that we use for our research are the ones that were approved by users.

Thanks to the data analysis, we can measure whether the newspaper is unbalanced in the direction of the public information needs, or if it tries to push a specific group of people more sensitive to a certain argument towards a determined discussion. The data could confirm or deny if the newspaper has succeeded in its purpose.

The dissimilarity between newspaper and social network can be assessed by defining a score measure, named “Points of consideration”. For example, if the newspaper push (score 8/10 given by the frequency and the visualization of the article on a definite topic on a precise date) on economic articles and (score 3/10) on justice articles, but the social network asks for (score 8/10 given by the frequency of people discussing on a particular hashtag or a content in a day) Justice articles and (score 3/10) economics articles, this idiosyncrasy can be detected within data correlating newspaper and social network. From the evaluation of people debating over a particular topic in a determinate moment on Twitter, it is possible to calculate the importance of it and its hashtag. This is what Twitter call “Trendtopic. From this principle it is possible to understand the importance of that subject. For newspapers side, it is possible to understand if an article catches the readers’ attention and give them the information by, the number of articles present in that newspaper, with how many people have viewed the articles on the same date about that topic. The equivalence of these two results gives the balance between the request for information in social media and the offer of

information from the newspapers, which will provoke the result of the ending the discussion on that trend topic.

Figure 1. An example of how it's possible to figure the "Points of consideration" between article and social media.

Figure 1 shows an example of connection between newspaper and social media data with some "points of consideration".

With a correlation approach, by studying the hashtags and articles shared, it is possible to understand who is acting as a micro-influencer on the social network and to connect it with those who want to augment/increase the popularity of the articles related to some specific argument.

Additionally, it is worth noting that the catalogue of hashtags is constantly growing. So, by moving back in the influence network, it is possible to understand who has shared the first hashtag (correlated to some news) and for which purpose. Also, this catalog will be helpful to understand the starting point of the news sharing process, for instance if some sudden or unexpected event has generated the evolution of a certain hashtag in time.

Moreover, it is important to know the entire set of those who "consume" the article as information on social media, in order to understand when this information has started to spread. Many newspapers can benefit from this service for the decision-making processes, allowing to grow some specific argument-sectors in the newspaper.

2.2. Hashtag and sharing information before Social media

Sometimes if the article is very old and keywords are not created as a hashtag, it could be difficult to understand when the topic started and when those articles became relevant on social media. Consequently, it is necessary to understand which article the editor promoted the most in the past, and how readers had used those articles on social media.

A data mining program can be used to understand article keywords and extract the possible hashtag associated to it. However, to choose the correct hashtag it is necessary a human-supervised action that include an examination of the cultural environment. For instance, the

political culture in the particular time frame when the articles were published could turn to be a crucial variable that can significantly modify the information in time.

As a consequence, a fundamental result of this project could be to understand the evolution of the newspaper by articles publication history. This could be done analyzing user's accounts that had shared the article in the past. Thus, it is possible to analyze the entire history section of the micro-influencers, i.e. the accounts that sponsored the post on Social media, but with some low kind of appearance in social media itself, and to see which person had previously shared the post on his profile.

The micro-influencer and the information analysis will be useful to understand why people on social networks had used some specific information to share ideas or to spread propaganda for his/her favorite candidate/party. By enlarging the perspective of the present work, the results of this analysis can show the evolution of the information over time. This could help other researchers to better understand the reasons for our current social and political condition, consequently the motivation behind certain political actions occurring in such a context.

3. Conclusion

Trend topics on social networks represent nowadays a valuable informative aid to newspaper, that gives the possibility to respond to unconscious requests for information (on a specific topic) on social network, with newspaper information supply.

Thanks to the use of big data, it is possible to evidence the unbalance of information request by people on social network and information supply by newspaper. This dissimilarity can be assessed by defining a score measure, named "Points of consideration", given by the elaboration of inner data from newspaper and patterns model from public data on social network. The equivalence of these two results gives the score balance between the request for information in social media and the offer of information from the newspapers, if in the end, the newspaper will supply correctly the information request with a specific high-quality article that will provoke the full information-satisfaction for people, as consequence there will be the end of the discussion on that trend topic.

References

Gallino Luciano, (2007) Tecnologia e Democrazia – *Biblioteca Einaudi*

Gustav Jakob Petersson & Jonathan D. Breul, (2019) Cyber society, big data, and evaluation. *Routledge editori*.

Lupton Deborah (2018). How do data come to matter? living and becoming with personal data. *Big Data & Society*.

Mazzeo Rinaldi Francesco & Giuffrida Giovanni, (2017) Big data e valutazione: una relazione ancora da costruire. *Rassegna italiana di valutazione* (aiv) n68 anno xxi issn 1826-0713, issne 1972-5027.

Resnyansky Lucy (2019). Conceptual frameworks for social and cultural big data analytics: answering the epistemological challenge. *Big Data & Society*.

Sadowsky Jathan (2019). When data is capital: datafication, accumulation, and extraction. *Big Data & Society*.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11638

Third Places and Art Spaces: Using Web Activity to Differentiate Cultural Dimensions of Entrepreneurship Across U.S. Regions

Timothy F. Slaper[1], Alyssa Bianco[2], Peter E. Lenz[2]
[1]Indiana Business Research Center, Indiana University, United States, [2]Dstillery, United States.

Abstract

We use unconventional, web-based user data to assess regional entrepreneurial activity and regional variations in characteristics and culture that drive differences in business formation. Using geographically granular, user-online activity to estimate a region's proclivity for entrepreneurship, we assess the statistical relationship between business formation, operationalized as establishment births, and a region's general interest in "third places" – informal gathering and mixing locations – and websites related to arts, music and design – "arts spaces." We operationalize interest in, or intention to patronize, third places and arts spaces by individuals within a geographical unit of analysis (U.S. counties) who access website information and resources related to those third places. Controlling for regional interest in entrepreneurship related web resources, we find that interest in third places and art spaces is strongly associated with regional variation in business formation. This work corroborates research showing that regions with a high concentration of interest (and participation) in third places and art spaces may attract the attention of would be entrepreneurs as desirable places to live, work and explore business opportunities, and help identify and address a critical missing ingredient in regions that have lower rates of start-ups and business growth.

Keywords: *entrepreneurship; business formation; website behavior; third places; regional culture.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

User-based data on website visits raise the possibility of enabling researchers and policy-makers to assess inter-regional differences in cultural attributes as they relate to economic performance, make economic data timelier and with greater geographic resolution.

The key contribution of this paper is to assess whether the web-user profile for interest in entrepreneurship, those interested in arts/design/music (ADM) and those visiting sites associated with third places (3P) like libraries and eating and drinking establishments are statistically related to business formation on a region by region basis. We make inferences that regional differences in these attributes may point to a population's culture as being more or less conducive to business formation and commercial innovation. We also confirm, or question, the degree to which various categories of ADM or 3P may have on business formation and innovation as often asserted by creative class and third places theories.

The paper is structured as follows. First, we highlight the touchstone literature relevant to our three interwoven strands of culture and entrepreneurship, nowcasting economic activity and 3P and ADM social theory. Second, we present our method and the unconventional data used. Third, we report the statistical results. Fourth, we conclude with a discussion about future research and possible policy applications.

2. State of the Discussion

Research on the potential of using the digital vapor trails, or digital exhaust, from tens of millions of users on tens of thousands of websites to predict economic outcomes, or nowcast economic activity, continues to grow. The often citied example is that of Glaeser et al. (2017) using online data sources, namely Yelp review data, to measure the critical economic activity of business formation well in advance of official government statistics. Other authors have joined the fray. Slaper and colleagues (2018a) linked interest in entrepreneurship related web resources with business formation for early start-up and growth phases of new ventures.

Two categories of regional attributes or characteristics may serve to fertilize the soil from which new businesses sprout: an arts and design concentration – often known as the creative class – and third places – meeting and mixing locations outside the home or work. Hawkins and Ryan (2013), for example, suggest that music festivals are a current form of third places for people and ideas to intermix and drive innovation. Creative class theory posits that there is a positive relationship between the arts and commercial innovation (Wojan and Nichols, 2018). Third places are also considered to provide a context and culture for the exchange of ideas and the coalescing of different types of talent needed to take an idea from the workbench to the market place (Oldenburg, 2001). Indeed, the Ewing Marion Kauffman Foundation developed the "1 Million Cups" initiative to provide space and resources

designed to help entrepreneurs discover solutions and engage with their local constituencies over "a million cups of coffee."

Finally, these three threads can be seen integrating and advancing regionally specific cultural (or personality) characteristics, unconventional data (from personality profiles) and business formation (or entrepreneurship) in the work of Obschonka et al. (2015), who aligned regional personality trait concentrations with business formation. Our study makes three small steps forward. One, using behavioral data to signal a region's cultural and personality profile. Two, using granular geographical units of analysis. Three, using data on user revealed preferences that is collected daily to (potentially) monitor and evaluate policy outcomes in real time.

3. Data and Method

We operationalize entrepreneurship as the count of establishment births by county, as captured and reported by the Census Bureau. An establishment birth – the opening of a new place of business – may not be a start-up, but rather the result of a multi-location corporation expanding its service region. However, this data source is the most accurate, consistent and of longest duration of publically available datasets. As control variables, we include county-based data for educational attainment and the Economic Research Service continuum of urban to rural counties – nine categories from large, highly populated urban counties to low population density counties distant from urban areas. We include a variable from the U.S. Internal Revenue Service: county-level data on the proportion of tax returns indicating individuals who made investments in non-IRA, self-funded retirement accounts (in contrast to company sponsored retirement accounts) as a possible signal for entrepreneurs and small business owners. The remaining data are derived from ZIP code based data from dstillery llc and aggregated to a county total. The data/variable selections are based upon availability of data collected by dstillery llc according to their client needs, but include ADM profiles such as arts avocation, design occupations, architecture & design occupations – variables used by Wojan & Nichols (2018). Those variables not previously studied are profiles of people interested in live music as well as other 3P such as libraries, golf courses, and casual dining restaurants. Categorizing live music is difficult as the venues attracting the web traffic may be considered 3P (Hawkins & Ryan, 2013), but the web traffic may also be those interested in the musical performance more than mixing with fellow audience members. Also included are the entrepreneurship web-resource user profile applied by Slaper and colleagues (2018a), as well as the previously unexplored venture capital occupation category and a small business resource profile. Lastly, two new potential 3P profiles/variables were included: one, "home brewers" who may have interest in micro-breweries and craft-beer third places; and two, volunteers at not-for-profit organizations. The latter can be strongly supported as both a third place for interpersonal mixing as well as an indicator of social capital.

The distillery llc is a marketing analytics firm that enables transactions between website hosts and the advertisers that pay to make web-based content free and accessible to the public. The distillery llc captures behavior on the web for their clients to post targeted, customer-specific advertisements. The distillery llc creates an analytical category/profile based on a bundle of websites that pertain to a particular constituency. In this case, the constituencies of interest are current or future entrepreneurs and those interested in 3P and ADM. The dstillery generated ZIP code-based user/device profile concentration data that we aggregated into county geographic boundaries based on population proportions and for which there are data on establishment births and the control variables. In all, of all the 3110 U.S. counties, 3035 counties had complete data for all variables. We performed an OLS regression to estimate the relationship between the explanatory and control variables and the log of the count of establishment births between 2014 and 2016. Except for the control variables and the IRS personal retirement investment proportions, all variables are indexes of website access concentration by county, with an index value of 1.0 being the U.S. national average. These index values are multiples of a region's interest concentration and website access behavior. For example, if Region A has an index for entrepreneurship resources of 2.0 while Region B has an index of 4.0, devices in Region B are twice as likely to access entrepreneurship resources as Region A devices and four times as likely as the average U.S. device.

The dstillery llc data are very timely, while the most recent establishment births data are from 2016. (The 2016 Census Bureau's Statistics of U.S. Business data is the most current at the time of this writing.) That said, cultural factors and regional personality traits, like population characteristics, do not change quickly over time. This follows Obschonka et al. (2015) both in terms of a region's personality profile helping to explain regional entrepreneurial and business formation activity as well as the fact that the psychological cultural characteristics of a region take multiple years if not decades to change.

4. Results

The dstillery llc consumer profiles are based on topical interests and as individuals access web resources for information and engage in economic transactions, their web behavior reveals their vocations, avocations, commitments and intentions for future purchases and activities. Each user profile is scored as an index value that is a multiple of the U.S. average of one. This allows easy comparisons within a user-profile and between profiles in the model. A coefficient of 0.23 for venture capital occupations indicates that a one-unit change in the index is associated with a 23 percent change in establishment births, while a coefficient of 0.64 for entrepreneurship resources indicates that for each one-unit increase there is an associated 64 percent change in establishment births. (As caveats: these interpretations are correct for only small changes, and the range and skew of indexes may differ dramatically from one user profile to the next. A big three U.S. auto-maker has a normal bell distribution with a maximum index of 5.8 while an office share company maximum index is above 60.)

Table 1: Regression Results of Third Places and Art Spaces Influence on Business Formation.

Dependent Variable, *county-based, Census Bureau*	**Log of Establishment Birth Counts 2014 to 2015 and 2015 to 2016**			
Explanatory Variables, *county-based, with data sources*	***Coefficients***	***Standard Error***	***t Stat***	***P-value***
Educational Attainment: population with Bachelor's Degree, percent – Census	0.03	0.00	7.45	0.00
Urban (1) to **Rural (9) continuum** – Economic Research Service (USDA)	(0.26)	0.01	(28.14)	0.00
Tax filings with non-IRA individual retirement investments, percent – IRS	37.05	2.62	14.17	0.00
Entrepreneurship Resources user profile concentration – d	0.64	0.14	4.62	0.00
Live Music Interest user profile concentration – d	0.43	0.06	7.64	0.00
Small Business Resources user profile concentration – d	0.61	0.09	6.60	0.00
Venture Capital Occupation user profile concentration – d	0.23	0.06	4.02	0.00
Library Interest user profile concentration – d	(0.53)	0.05	(11.64)	0.00
TGIFridays casual dining user profile concentration – d	0.10	0.03	3.77	0.00
Buffalo Wild Wings casual dining user profile concentration – d	0.03	0.03	1.09	0.28
Golf Course venues user profile concentration – d	0.02	0.01	1.95	0.05
Arts Avocation user profile concentration – d	(0.44)	0.10	(4.17)	0.00
Design Applications user profile concentration – d	0.58	0.09	6.22	0.00
Architecture & Design user profile concentration – d	(0.14)	0.07	(2.07)	0.04
Volunteers serving in Non-Profit user profile concentration – d	0.01	0.08	0.13	0.90
Home Brewing user profile concentration – d	(0.56)	0.05	(11.51)	0.00
Intercept	4.80	0.10	47.91	-
Adjusted R Square	**0.73**			
	df	***SS***	***MS***	***F***
Regression	16.00	5,291.96	330.75	513.43
Residual	3,036.00	1,955.78	0.64	
Observations	3,053			

Source: d = dstillery llc.

The selected profile variables to represent ADM, 3P and entrepreneurship resource seekers were statistically robust. As reported in **Table 1**, the adjusted R-square was 0.73. Only the variables for the Buffalo Wild Wings, in contrast to the other casual dining third place, and non-profit volunteers were not statistically significant at 0.05 level. There were surprises, however, as several of the coefficients' direction and magnitude were in the opposite direction of expectations. Coefficients for home brewers, library, arts avocation, and architecture & design profiles were negative. The home brewer profile, it appears, does not well represent those frequenting micro-brewery drinking establishments to socially mix but rather those aspiring to brew the craft beers served as such establishments. Libraries do not appear to be 3P for those starting new firms or older firms expanding their geographic reach. Considering the activities associated with libraries – finding information but not interacting and mixing with other patrons – absent special programs hosted at a library, these institutions are not well-suited for the exchange of ideas or making business contacts.

The more puzzling negative results are associated with arts avocation and architecture and design occupation profiles. It may well be that the live music profile is capturing the same statistical association with ADM as found by Slaper and colleagues (2018b) and Wojan and Nichols (2018). Live music performances may be reasonably categorized as either a third place or an ADM art space. Indeed, the profiles for arts avocation would include musicians and other performers, as well as artists and sculptors. Together, the magnitude and sign of the coefficients for arts avocation, design applications and architecture and design occupations tend to cancel one another, raising questions about whether third places and art spaces are complementary or whether one influence dominates.

The other statistical associations are in line with expectations: a positive relationship between educational achievement and business formation as well as the negative coefficient for the urban to rural spectrum that runs from one (dense-urban) to nine (sparse-rural). One would expect the rate of business formation to be scale dependent (urban), with greater populations, density and Jacobian interactions fostering higher rates of business formation together with higher levels of educational achievement.

5. Conclusion

We find support for the importance of third places and art spaces – as measured by a region's interest and, by extension, intention to participate in these pursuits – on business formation. Our study corroborates with the work of Obschonka and colleagues (2015) on the importance of regional culture on business dynamics, broadens the application of new sources of data to measure and monitor economic activity in almost real time in a manner similar to Glaeser and colleagues (2017), and strengthens the relevance of social theory regarding third places a'la Hawkins and Ryan (2013) and the creative class, e.g., McGranahan and Wojan (2007).

These encouraging results warrant further exploration into developing sets of variables that more rigorously correspond with theories of economic and social outcomes as they relate to preferences and intentions as expressed by web behavior. Differentiating the beneficial influences of 3P in contrast to ADM would help secure the position of one theory over the other in terms of the strength of the effect on the exchange and synthesis of ideas and knowledge spillovers. Are the confluence venues and magnitude of social mixing more salient factors than the concentration of the creative class? Are there other regional cultural contours that may be currently unobserved that may reveal other important regional traits?

These unconventional and timely data have great promise to now-cast current economic activity or presage economic outcomes. These results, albeit early, show that there are regional cultural differences that may explain divergent outcomes of policies and initiatives designed to promote innovation and entrepreneurship. Researchers may be one step closer to identifying the scarce ingredient in a region's cultural context that constrains business dynamism. Supplementing that scarce ingredient would be a worthy pursuit for policymakers to encourage innovation, business dynamics, economic growth and social well-being.

References

Elmborg, J. K. (2011). Libraries as the spaces between us: Recognizing and valuing the third space. *Reference & User Services Quarterly*, 338-350.

Glaeser, E. L., Kim, H., & Luca, M. (2017). *Nowcasting the local economy: Using yelp data to measure economic activity* (No. w24010). National Bureau of Economic Research.

Hawkins, C. J., & Ryan, L. A. J. (2013). Festival spaces as third places. *Journal of place management and development*.

McGranahan, D., & Wojan, T. (2007). Recasting the creative class to examine growth processes in rural and urban counties. *Regional studies*, *41*(2), 197-216.

Obschonka, M., Stuetzer, M., Gosling, S. D., Rentfrow, P. J., Lamb, M. E., Potter, J., & Audretsch, D. B. (2015). Entrepreneurial regions: do macro-psychological cultural characteristics of regions help solve the "knowledge paradox" of economics?. *PloS one*, *10*(6).

Oldenburg, R. (Ed.). (2001). *Celebrating the third place: Inspiring stories about the great good places at the heart of our communities*. Da Capo Press.

Slaper, T., Bianco, A., & Lenz, P. (2018, September). Digital Vapor Trails: Using Website Behavior to Nowcast Entrepreneurial Activity. In *2nd International Conference on Advanced Reserach Methods and Analytics (CARMA 2018)* (pp. 107-113). Editorial Universitat Politècnica de València.

Slaper, T., Wojan, T., Crown, D., & Lenz, P. (2018b, November). Are the Problem Spaces of Economic Actors Increasingly Virtual? What Geo-located Web Activity Might Tell Us about Economic Dynamism. Presented at the North American Regional Science Council Annual Conference, San Antonio, TX.

Wojan, T. R., & Nichols, B. (2018). Design, innovation, and rural creative places: Are the arts the cherry on top, or the secret sauce?. *PloS one*, *13*(2).

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11642

New technologies and role of direct surveys in the production of Official Statistics

Loredana De Gaetano, Pasquale Papa
ISTAT – Italian National Statistical Institute, Rome, Italy.

Abstract

The technological development applied to statistical data collection processes, that has taken place in recent years, and the increasing availability of alternative statistical sources, notably administrative and so-called "sensor", is leading to a profound revision of the role of direct surveys, also in the context of Official Statistics.

The objective of this article, based on the experiences in progress in ISTAT (Italian National Statistical Institute) is to evaluate the status of the transition towards a more modern, efficient and sustainable way of conducting direct surveys, in a set of specific areas, providing examples and targeted analysis. This analysis will help to build a general framework towards which the collection of Official Statistics data will converge in the next years.

In the above mentioned framework the analysis will mainly involve three converging macro areas: a) individuation of efficient management set-up of data collection processes; b) application of innovative techniques in the data collection of traditional direct surveys; c) aspects related to the availability of new alternative sources to those currently used in the production of official statistics.

Keywords: *smart surveys; data collection; centralised data collection; administrative sources; sensor data; data collection statistical portals*.

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

The opportunities offered by the application of new technologies to the field implementation of direct surveys, that has taken place in recent years and the increasing availability of alternative statistical sources mainly in the form of administrative archives, and other sources not directly produced for statistical purposes is leading to a profound revision of the role of direct surveys, also in the context of Official Statistics.

At the same time, various trends have emerged in the social and economic contexts of many countries which tend to reduce the participation of the statistical units involved in the surveys, recording average decreasing response rates. The awareness of the increasingly less sustainable statistical burden that official statistical surveys exercise on respondents has also developed, showing the growing need to reduce it.

2. Investigated areas

The effects of all these trends led to the development of new strategies in the management of data collection in Official Statistical processes, summarized in the following thematic areas:

2.1. Data collection processes management set-up

The first step consists of adopting management solutions that allow greater control over the data collection processes conducted in the context of Official Statistics. An example in this regard is the adoption of an integrated and centralized data collection approach. This solution involves greater efficiency of data collection procedures and control on total direct survey quality. A second step concerns the development of generalized survey management systems, as much as possible oriented towards integration and rationalization that allow making all the activities required to the various actors involved in the survey data collection processes more coordinated and efficient. The adoption of multi-technical approaches in order to provide tools that adapt to the different types of users involved in the surveys and their structural characteristics (age and cultural level for individuals and size and sector for companies) is also very important.

2.2. Application of advanced techonologies for data collection purposes

Technical development allows the application of advanced technologies to direct survey data collection according to an approach that can be defined as "smart". At the same time in the last years several National Statistical Institutes have started designing “Portals” for data collection. Development of data collection Portals aimed at optimizing relations with respondents and providing specific services aimed at rationalizing and simplifying the statistical requirements. These portals are particularly effective to encourage the development of the CAWI technique which allows respondents to fulfill the statistical requirements by

filling in web questionnaires. Portals also involve the development of a series of user-oriented facilities aimed at simplifying the task required of the respondent (e.g. centralized inbound and outbound Contact Center services). ISTAT (Italian National Statistical Institute) implemented a portal for data collection to all companies involved in economic surveys starting from 2015 (*Portale Statistico delle imprese*). It involved the exploitation of the CAWI technique, reducing the costs for the remuneration of esternal companies used for the management of CATI and CAPI techniques. A single portal is currently being designed for the units involved in all direct surveys. A theme on which Istat is also investing is that of the use of artificial intelligence to manage some repetitive phases of the data collection process, e.g. automated sorting requests for assistance coming from the units involved in the surveys to the correct recipients and automatic response to repetitive questions coming from respondents.

2.3. Use of alternative sources for statistical purposes

The development of alternative sources mainly consists of administrative data that can be used for statistical purposes. In this regard, ISTAT has launched two parallel initiatives: the first consists in the design of an integrated system of registers, based mainly on administrative sources, in order to support and complement the direct survey-based data collection. The second initiative concerns the use of administrative archives for the specific needs of specific direct surveys. All this requires the development of synergies between entities producing official statistics or owners of administrative sources used for statistical purposes. In addition to administrative sources, other data sources can be used which are already available for different purposes: e.g. big data and sensor data. In this regard, Istat is planning specific techniques for retrieving information from these data sources.

3. Conclusions

The main trends in the field of data collection for Official Statistics, which involves management, technical and contextual aspects imply a radical revision of the role of direct surveys. In fact the role of direct surveys tends more and more to be limited to sectors and situations where alternative sources are not available or to measurement of distortions of the alternative sources themselves. In this perspective, the direct surveys must become smaller and more targeted and qualitative, and the need to produce instruments aimed at measuring and controlling the Total Survey Error (TSE) which includes both the sample and non-sample components of the statistical error still remains a topic of great interest. In the above mentioned framework the objective of this article, based on the experiences in progress in ISTAT, is to evaluate the progress of the transition in each of the identified areas, providing examples and targeted analysis. This analysis will help build a general framework towards which the collection of Official Statistics data will converge in the next years. The transition

also tends to meet two basic requirements for Official Statistics of the future, that is reducing costs, with a view to promoting the social sustainability of official statistical systems.

References

Bellini G., Monetti F., Papa P. (2018), The impact of a centralized data collection approach on response rates of economic surveys and data quality: the Istat experience. Q2018, European conference on quality in official statistics, Kraków

Bellini G., Cecconi N., De Gaetano L., Monetti F., Papa P. Ranaldi R. (2018),Centralizing data collection implementation: the Istat experience: Data Collection Workshop 'Resourceful Data Acquisition- UNECE Geneva

De Gaetano L., Digrandi A., Papa P. (2018), A territorial model for data collection implementation. Q2018, European conference on quality in official statistics, Kraków

Fazio N.R, Murgia M., Nunnari A. (2013), The business statistical portal: a new way of organizing and managing data collection processes for business surveys in Istat, Unece - Conference of european statisticians.

Istat (2017), Mapping delle attività della DCRD nell'ambito dello schema concettuale di riferimento internazionale GSBPM. Delibera D16 49 DIRM2017.

Istat (2016), Istat's modernisation programme,

https://www.istat.it/it/files//2011/04/IstatsModernistionProgramme_EN.pdf

Lise Rivais, Marc St-Denis, Susan Lensen (2013), Centralising data collection at Statistics Canada. Seminar on Statistical data collection. Unece - Conference of european statisticians.

Saraiva dos Santos P. and Moreira A. (2013). Creating a data collection department: statistics portugal's experience. Seminar on Statistical data collection. Unece - Conference of european statisticians.

Signore M. (2017) GSBPM and other international standards MedStat training on GSBPM, Istat - Rome (2017-07). Available at https://statswiki.unece.org/display/GSBPM/GSBPM+Training+Materials

Signore M. (2017) GSBPM how to use and implement MedStat training on GSBPM, Istat - Rome (2017-07). Available at https://statswiki.unece.org/display/GSBPM/GSBPM+Training+Materials

Marske R. And Stempowski D. M. (2009), Company-Centric Communication Approaches for Business Survey Response Management, Statistics Canada Symposium 2008: Data Collection: Challenges, Achievements and New Directions. Available at https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2008000/article/10983-eng.pdf?st=0-IwAKvz

Unece Statistics wiki. Generic Statistical Business Process Model – GSBPM. http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

Groves, Robert, Floyd J Fowler, Mick Couper, James Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. Survey Methodology. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Groves R.M. and Lyberg L. (2010) Total Survey Error: Past, Present, and Future *Public Opinion Quarterly*, Volume 74, Issue 5, 2010, Pages 849–879, https://doi.org/10.1093/poq/nfq065

Biemer P. Total survey error design, implementation, and evaluation Public Opinion Quarterly, Volume 74, Issue 5, 2010, Pages 849–879.

Geon Lee G., Benoit-Bryan J. and Johnson T. P. (2012), Survey Research in Public Administration: Assessing Mainstream Journals with a Total Survey Error Framework, Public Administration Review, Vol. 72, No. 1 (JANUARY/FEBRUARY 2012), pp. 87-97

Jedinger A., Watteler O. and Förster A. (2018), Improving the Quality of Survey Data Documentation: A Total Survey Error Perspective- *Data* — Open Access Journal

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11643

Sample Size Sensitivity in Descriptive Baseball Statistics

John Kulas[1], Marlee Wanamaker[1], Diuky Padron-Marrero[1], Hui Xu[2]
[1]Department of Psychology, Montclair State University, Montclair, NJ, USA, [2]US Bancorp, Minneapolis, MN.

Abstract

This paper presents one element of a larger project that probes for systematic and predictable patterns of variability/volatility in baseball's descriptive statistics. The larger project standardizes many baseball indices along an event metric and provides relative estimates of each index's point of inflection toward an empirical asymptote. Specifically these estimates reflect deviations in sensitivity to "sample size" (e.g., which descriptive statistics are more or less robust across events). The end purpose of this broader investigation is a qualifier to be associated with such statistics: sample size sensitivity (Triple S). Not because it's needed, but because, colloquially, discussions of baseball statistics are commonly qualified by the cautionary statement, "well, it's a small sample size". The current presentation highlights the process and results of estimating the logarithmic event function of one statistic, batting average, and we will provide real-time projections of accuracy (our estimated function versus in-coming baseball data that occurs during the CARMA conference). Results have implications for the integration of BigData applications into digestable summary statistics that appeal to a broad-reaching audience with practical implications and meaning.

Keywords: *logarithmic function estimation; baseball data; predictive model.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

Job performance is dynamic – e.g., it changes over time and context (see, for example, Sturman, 2003). Although the dynamic nature of performance has been acknowledged for a very long time, Kane (1996) was perhaps the first to propose that researchers should conceptualize job performance as a *distribution of outcomes*. Within the social sciences, this perspective was originally cited as impactful, but as of February 2020, this unique conceptualization of worker performance had only realized 76 academic citations. Recently, however, the potential of BigData to inform all aspects of work (including performance) has been met with a proliferation of interest and investigations (e.g., Campion, Campion, & Campion, 2018; Gunasekaran, Papadopoulos, Dubey, Wamba, Childe, Hazen, & Akter, 2017; Tonidandel, King, & Cortina, 2016), including the potential to revisit Kane's (1996) proposal of defining job performance via dynamic functional distribution.

Parallel to the emergent interest of BigData applications to organizational phenomena, in October of 2017 the Journal of Business and Psychology published a special issue dedicated to the interdisciplinary relevance of athletics and organizations. In their initial call for paper and subsequent introduction, the special issue editors noted "how studies of sports can readily be compared to and applied to the study and practice of work in organizations" (Gentry, Hoffman, & Lyons, 2017, p. 509). The current CARMA presentation integrates these two relative "newcomers" into the organizational sciences: athletics and BigData. Specifically, we resurrect Kane's (1996) perspective on performance distributions, leveraging baseball data to inform the modeling of performance over time.

Dalal, Nolan, and Gannon (2017) posed similar questions, tracking performance (goals, assists, and positive/negative differential) based on the occurrence or absence of previous shared experience with teammates (their sample was Olympic hockey players, permitting an estimate of players who had and hadn't previously been "teammates"). They noted the particular relevance of their sample to the construction and utilization of temporary teams used by traditional corporate organizations.

Similarly oriented, Heazlewood (2006) attempted to predict performance of Olympic swimming athletes. He found that nonlinear models were better predictors of performance than linear models. Results were also more accurate for races that were shorter distances. The predictions were made using mathematical models that predicted performance in 1996 and 1998 and were evaluated based on how closely they predicted performance in events that had not happened yet at the time of the analysis. These nonlinear models were again noted as patterns also likely to occur within more traditional worker contexts – perhaps having implications on the duration of work tasks and suggesting qualitatively different approaches toward modeling performance across different task periods.

Hofmann, Jacobs, and Gerras (1992) applied a historical equivalent to our current pursuit: "mapping" performance across time in two samples of baseball players. Their interest was in relative rank orderings scross time and the stability of such orderings. Due to data capturing limitations of the age, these authors were reliant on annual summary data from 204 professional baseball players presented within *The Baseball Encyclopedia* (1990). Their findings suggest a common nonlinear inverted-U shaped trajectory for offensive performance (batting average), with pitching data (earned run averages) exhibiting more linearity over time (e.g., ERAs deteriorated fairly consistently across years played). They note possible implications regarding patterns of performance for traditional workers across seniority and tenure.

The current CARMA presentation utilizes similar information as Hofmann et al. (1992), but does so with the advantage of contemporary data-capturing capabilities. Specifically, we capture *event-level* data (e.g., each pitch of a baseball) in an attempt to model differences across descriptive statistic stability. For the current presentation, we focus on one index of offensive performance: batting average. Across players and years, we model the functional degradation and eventual stability of this statistic, and use this empirical function to predict player performance during the July 8-9 conference period.

2. Methods

Play-by-play data from all regular season major league baseball games played from April 2008 to October 2015 was retrieved from baseballsavant via Bill Petti's database building script (https://billpetti.github.io/2018-02-19-build-statcast-database-rstats/). Each datafile contains approximately 700,000 individual *plays* – the most common form of play is a *pitch* (that is, the pitcher throws the baseball to his catcher, while a batter either attempts to swing or not). For the purposes of batting average, we collapsed these individual pitching plays into offensive player *at bats*. An *at bat* is a plate appearance that results in our focal event – the presence (1) or absence (0) of a "hit".

Each year, every offensive player begins with a simple batting average of zero. After one *at bat* the player's batting average either stays at zero (he did not record a hit) or rises to 1.0 (he did record a hit of some sort – a single, double, triple, or home run). Upon subsequent *at bats*, the player's batting average reflects the cumulative number of hits divided by the cumulative number of opportunities (at-bats). Eventually, most batting averages tend to stabalize due to the sheer number of opportunities accumulated throughout the baseball season (the denominator of the batting average statistic becomes quite large, effectively neurtering the influence of the binary numerator event [hit (1) or miss (0)].

After computing sequential cumulative batting averages for each player across the course of his full season, we next calculated absolute batting average difference between each player's

at bat. Figure 1 illustrates this information with a small subset of 2008 data - these are American League (AL) first basemen. The x-axis reflects the *number of at-bats* and is truncated at 160 for simplicity of visual presentation (e.g., our goal was to make the Figure 1 presentation as easily interpretable as possible). Here, the x-axis origin reflects the progression from a player's *first* at bat to their *second* (because each player's first at bat was the first meaningful recording of the possible hit event). The y-axis reflects the absolute deviation from the first to second event – as can be seen in Figure 1, the largest absolute deviation from the first to second at bat is .5 – this happens when the hitter alternates event outcomes across the two at bats (e.g., misses the first and hits the second or vice versa). This left-most event also represents the greatest opportunity for a large index due to the small denominator (2).

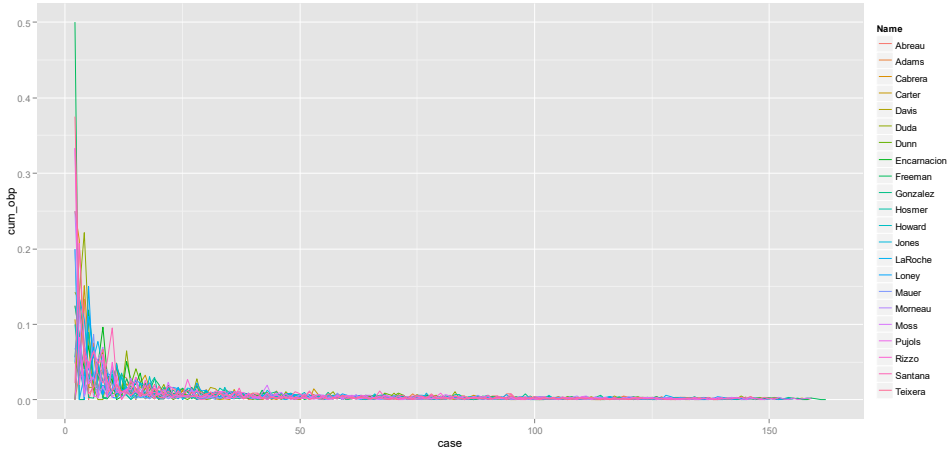


Figure 1. Small subset example of raw cumulative (absolute) average discrepancy.

3. Results

For purposes of function estimation, we were interested in the variance within vertical arrays. Figure 1 reflects this systematic pattern of heteroskedasticity, with greater variability in estimates near the origin (and less variability as plate appearances increase; e.g., to the "right" of Figure 1). The pattern is a bit more visually evident with the *standard deviation* of average absolute discrepancies, and these are presented for the first 100 at bats across all offensive players for the 2008 baseball season (see Figure 2).

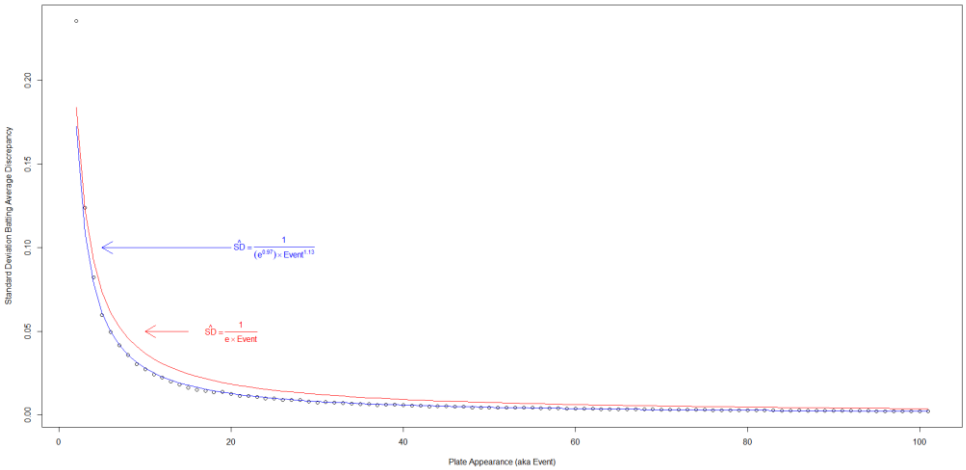


Figure 2. Predicted standard deviation as a function of event (in this case, "at bat" aka plate appearance).

Fitting a logarithmic regression to these standard deviations yields a reasonably predictive function: the predicted standard deviation (within vertical array) approximates $\frac{1}{e*Event}$. Via application of 8 years of baseball data, however, we were able to specify very slight modifyers to this general pattern: our empirical regression equation has slightly modified intercept and slope. For example, the 2008 data function was: $\log\left(\widehat{\frac{1}{sd}}\right) = .97 + 1.13 * \log(event)$. These functions explained the patterns of heteroskedasticity very well (R^2 = .9975, F = 39,690, $p < .05$ [again, only 2008 data]), and is presented visually via the blue function in Figure 2. Algebraically, our predictive model (solving for standard deviation in plate appearance batting averages instead of a logarithmic transformation of these) simplifies to: $\widehat{sd} = \frac{1}{e^{.97} * event^{1.13}}$. We also estimated similar functions for the other seven years of retrieved data. The CARMA presentation is dynamic, updating MLB player events and presenting as residual values to our aggregated (across 8 years) predictive function for batting average stability.

4. Discussion

For the purposes of this presentation, we focused on modeling the heteroskedasticity of a descriptive baseball statistic via standard deviation specification – by computing a simple standard deviation within each "array" (arrays are performance events – in Figure 1 the x-axis represents these events [e.g., an MLB "plate appearance"]). In broader applications, an "event" can be a "widget" (e.g., production), a work period (e.g., hour, shift, week, month),

or service event (e.g., customer/consumer rating). Our ultimate interest, therefore, is twofold: 1) we intend to model similar functions across different baseball statistics, taking note of functional asymptotes and points of inflection, and 2) we hope to apply the general procedure of functional estimation across events to more common occurences of performance. Sturman (2003) notes in his meta-analysis that performance trends across time do tend to be different for different types of job, and so the estimate of functions across different baseball indicies may very well parallel different functions estimated across different jobs. Similar to the perspectives of both Kane (1996) and Hofmann, Jacobs, and Gerras (1992), our ultimate goal is to leverage insights taken from athletic performance in an attempt to conceptualize job performance in a new manner (here being operationalized as a predictive function across events).

References

Campion, M. C., Campion, M. A., & Campion, E. D. (2018). Big data techniques and talent management: Recommendations for organizations and a research agenda for IO Psychologists. *Industrial and Organizational Psychology*, *11*(2), 250-257.

Dalal, D. K., Nolan, K. P., & Gannon, L. E. (2017). Are pre-assembly shared work experiences useful for temporary-team assembly decisions? A study of Olympic ice hockey team composition. *Journal of Business and Psychology, 32,* 561-574.

Gentry, W. A., Hoffman, B. J., & Lyons, B. D. (2017). Box scores and bottom lines: Sports data can inform research and practice in organizations. *Journal of Business and Psychology, 32,* 509-512.

Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S. F., Childe, S. J., Hazen, B., & Akter, S. (2017). Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research*, *70*, 308-317.

Heazlewood, T. (2006). Prediction versus reality: The use of mathematical models to predict elite performance in swimming and athletics at the Olympic games. *Journal of Sports Science and Medicine, 5,* 541-547.

Hofmann, D. A., Jacobs, R., & Gerras, S. J. (1992). Mapping individual performance over time. *Journal of Applied Psychology*, *77*, 185-195.

Kane, J. S. (1996). The conceptualization and representation of total performance effectiveness. *Human Resource Management Review, 6,* 123-145.

Ployhart, R. E., & Hakel, M. D. (1998). The substantive nature of performance variability: Predicting interindividual differences in intraindividual performance. *Personnel Psychology, 51,* 859-901.

Sturman, M. C. (2003). Searching for the inverted U-shaped relationship between time and performance: Meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management, 29,* 609-640.

The baseball encyclopedia (8th edition). (1990). New York: MacMillan.

Tonidandel, S., King, E. B., & Cortina, J. M. (2016). Big Data methods: Leveraging modern data analytic techniques to build organizational science. *Organizatinoal Research Methods, 21,* 525-547.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11645

Extracting usual service prices from public contracts

Tomáš Bruckner, Filip Vencovský
Department of Information Technologies, University of Economics Prague, Czech republic.

Abstract

The paper describes a project of automatic selection, scraping, and full-text analysis of contracts in the area of IT and Information Systems. The purpose of the project was to extract manday prices and build the list of usual manday prices for particular roles that are stated in the contracts. The list aims to provide a foundation for sizing of new IT solutions before the public tender for an association of major state institutions of the Czech Republic. The result of the research is the list of usual prices for the specified roles, including blended rate, based on median and interval between quartiles, all with demonstrable links to origin contracts. The discussion states additional social factors to be considered when interpreting and using the resulting list, like the subjective influence of validators, tendency for generalization, or defensive attitude of affected vendors.

Keywords: *usual price; contracting; full text analytics; information technology, big data.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Usual prices are used in businesses to set the estimated value of procurement. Especially IT or Information Systems contracts tendered by public institutions are sensitive for correct sizing (Ochrana & Pavel, 2013) due to the special rules of public tenders where preliminary negotiations for the purpose of finding expected price can be challenged by competitors who were not addressed. Public institutions struggle with the need of usual price determination before any public tender. Therefore, thirteen IT departments of major state institutions joined in an association with the purpose of tenders facilitation and assigned and sponsored this research.

The usual prices are generally obtained by a counsellor, who takes few demonstrably negotiated contracts and proves prices negotiated in a given time and place, as stated in (Act. No. 526/1990). In IT and Information Systems contracts, such findings can vary greatly and hence are not very useful for contract sizing. Due to that, the sponsor decided to order research of broad usual prices overview, based on all contracts demonstrably negotiated by public institutions in the Czech Republic, which are available in the public contract register (Act No. 340/2015).

The research problem is a complex application of technology and business topics. The technology consists of web scraping, unstructured data analysis, information extraction, information quality and data validation, business intelligence and data visualization. The business part consists of understanding the content and principles of contracting and pricing in IT and Information Systems. The primary source of the data is a public register of contracts, available on the internet, which contains millions of contracts. The problem is to restrict the amount to those contracts, which are relevant to IT and Information Systems, read the contracts, identify the place and time of validity, identify price details, extract the price details, classify the pricing data, decide the right usual price and visualize and present the data in the form beneficial for the sponsor institution employees on a daily basis.

The level of correctness has to be high because of the result is observed and verified by the potential vendors of the future contracts. The vendors felt offended in the preliminary stage of the research; arguing the projects dictates the contract prices and restricts freedom of price setting and giving the sponsors an overview about different prices the same vendor contracted with different state institutions. Therefore, the communication of the research and its results must be very accurate and cautious.

The prices of contracts in IT and Information Systems are complex pricing systems assembled of unit prices. The unit prices could be the prices of services, such as licenses, access to applications, computing time, or products like hardware. This paper focuses on a specific unit price - a manday price of work of a specialist.

2. Methodology

We decided the main measure of the result is correctness, where every price included in the final result should be traceable to the original contract. Thus, every possible protest against the result could be argued easily. The proposed list of the usual prices shows typical roles of specialists in IT or Information Systems area. The usual price was calculated for each of the roles as an interval based on a set of traceable data extracted from the contracts. The additional value for users of the list was provided by deeper classification of each role using tags. Hence, the price can be estimated more precisely for various conditions. The list of usual prices was accompanied by a benchmark of prices negotiated by the sponsor institutions. The initial role set, a role codebook, was interviewed with future users of the list within sponsor institutions.

As the prices may change in time, the list should be repeatedly assembled in a certain interval. We set the interval to six months in which the two years of historical data will be evaluated. Due to the repeatability, the research method should be as automated as possible and independent on individual decisions of the researches of future runs but should enable partial innovations of the process. We elaborated the documented research process, consisting of sub-processes: Preparation, Contract acquisition, Prices extraction and validation, Usual price determination, Handover and acceptation.

In the preparation phase, besides the infrastructure preparation (the technological architecture composed for the research is the subject of another paper (Bruckner, 2019)), a list of organizations operating in particular business areas concerning IT and Information Systems is created. The reason of the list is to exclude contracts, which are surely not related to the area in question. For that purpose, the NACE activity classification (Nomenclature statistique des activités économiques dans la Communauté européenne) from the Public Business registry is taken into account.

The contract acquisition phase is performed as a combination of scraping the frontend of the public register of contracts and API access to the register. Only contracts of subjects listed in the previous step by NACE are scraped. The scraping resulted in the set of contracts that were downloaded in the next step. Each downloaded contract is transformed into plain text, whereas the original form is preserved for a possible visual check. In some cases, OCR over an image must be performed. All contracts in plain text are reverse-indexed, lemmatized, stemmed and stored in the search engine (Zobel & Moffat, 2006) . As the search engine, we employed Elasticsearch (“Elastic Enterprise Search,” 2020), a software that uses Apache Lucene search library. Every half of a year a new run is indexed into a separate database. Every contract is identified by a specific number, new version of the same contract rewrites older version.

The prices extraction phase consists of document indexing, search, and processing. As the contracts were indexed, we were able to create sets of queries for the selected typical roles of specialists in IT or Information Systems area to identify contract documents, in which a manday price for those roles probably occurs. The results of the queries were experimentally verified by experts.

The queries were executed on the search engine. The documents were processed by price extraction scripts in the next step. The price extraction is realized by a set of regular expressions (Brauer, Rieger, Mocan, & Barczynski, 2011; Mooney & Bunescu, 2005) which match several possible ways to describe a manday price in the text of the contract and which most likely do not match other data than prices, e.g. quantity of mandays etc. The regular expressions were manually verified on a subset of data and were a subject of fine-tuning during the price extraction.

In the validation phase, the group of experts verified the correctness of extraction of every price data. The extreme or unusual data were validated by two experts at least. Data for validation were sorted ordered according to their relevance based on the TF-IDF measure (Salton & Buckley, 1988) in a large table, with several contract metadata, and submitted to a group of experts for validation.

In this paper, we focus on the method of determination of the resulting list of usual prices. Thus, the core activity took place in the usual price determination phase. The handover and acceptation phases are not described in this paper.

We define the usual price as a price for which, in given time and given place, was provided (sold) a similar service or product. The given time, we understand as the period of two following years for which we gathered the data. The place, in this case, is the territory of the Czech Republic.

By a similar service, we understand a provision of a service priced by a rate for a time interval (an hour or a day) when providing service in the context of information technologies or information systems. For the classification of a particular service to one of the predefined roles, the description of the service in the contract is determinative. We reserved the right to classify hazily or shortly defined service descriptions inaccurately.

All price data were normalized to a manday price, a price for eight hours of work, in local CZK currency, VAT exclusive. If the currency is not CZK, the price is converted by Czech CEB conversion rate valid at the day of signing the contract.

Due to the variability, we do not determine the usual price as a price of single services but as a statistical calculation from a set of single particular services in given time and place for every class of typical role. From the set, we compute the following descriptive statistics: lower quartile, median, upper quartile, minimum, average, and maximum.

For the determination of the usual price, we suggest not to take into account the extreme values. Thus, we consider the usual price as the median. Moreover, we define a usual price spread as the spread between quartiles, or the interval, which do not include a quarter of lowest and a quarter of highest prices within the given role set.

Due to the method of data acquisition, prices for the same role from one contract were obtained multiple times. There were cases in which the stated price differs for the same roles; usually due to the different time of negotiation. Therefore, if the set includes different prices for the same role classified by the role codebook and tag description from the same contract, such prices were included only once, namely as an average.

If there were stated different price data for different roles in one contract, all of them were included in the data table separately. The same rule was applied in case of different specification of the role was detected, e.g. different seniority or request response time. The different specification data were included in the data table with the use of corresponding tags.

Considering the existence of various tags, we included tag consolidation activity into the usual price determination phase. In the time of extraction and validation, we created a semantic mapping of different tags. This mapping was used to join small sets of differently tagged price data to bigger, more relevant sets. In this step, the size of the set and the semantic generalization of the tags were considered. The consolidation was done by regular expressions with several variants of tag generalization.

The final output was a set of tables: summary table of all characteristics for roles, summary histograms for particular roles, and summary box plot chart in the same scale for roles. Furthermore, for every role was created a separate table where, apart from summary values for the role, values for specific subsets defined by consolidated tags were presented.

3. Results

We completed the list according to the described method five times, for the first time in half of 2017, for the last time in the half of 2019. The last run contained the contract data from May 2017 to May 2019. The method has undergone some slight innovation on the way. The method and the results presented in this paper are from the last run.

The complete set of all contracts in the public contract register counted millions of documents. After the NACE restrictions, we scraped amount of 566,315 contracts. After the price extraction step, we got a set of 10,051 candidate price data items. After manual validation and tags consolidation, we finished data preparation with 6,408 price items from 2,336 unique contracts from 776 public contracting authorities and 884 vendors.

Table 1. Summary of usual prices for roles.

Role	Lower quart	Median	Upper quart	Min.	Avg.	Max.	Con-tracts	Ven-dors	Pri-ces
administrator	5,920	7,600	10,560	1,200	8,567	24,000	213	133	357
analyst	8,000	10,400	13,580	1,440	10,701	24,000	341	149	402
architect	7,422	11,920	12,800	1,500	10,906	42,368	140	55	200
auditor	9,600	11,400	14,830	6,400	12,079	20,400	18	18	19
blended_rate	7,376	9,600	12,000	800	10,099	47,120	559	285	671
support	6,000	8,000	12,000	8	9,364	62,400	521	283	757
coder	7,260	9,600	11,940	1,920	9,697	30,560	553	242	652
routine_jobs	4,720	8,000	10,400	1,512	8,297	24,000	174	90	238
specialist	8,000	10,000	12,000	1,040	10,689	42,368	849	336	1,650
lecturer	6,400	9,600	12,000	1,800	9,952	43,200	360	167	433
technic	4,590	6,400	8,650	40	7,122	31,680	380	289	612
tester	5,680	7,912	9,975	1,232	8,038	20,584	105	62	123
project_manager	8,160	12,000	14,400	1,300	11,852	28,000	236	104	294

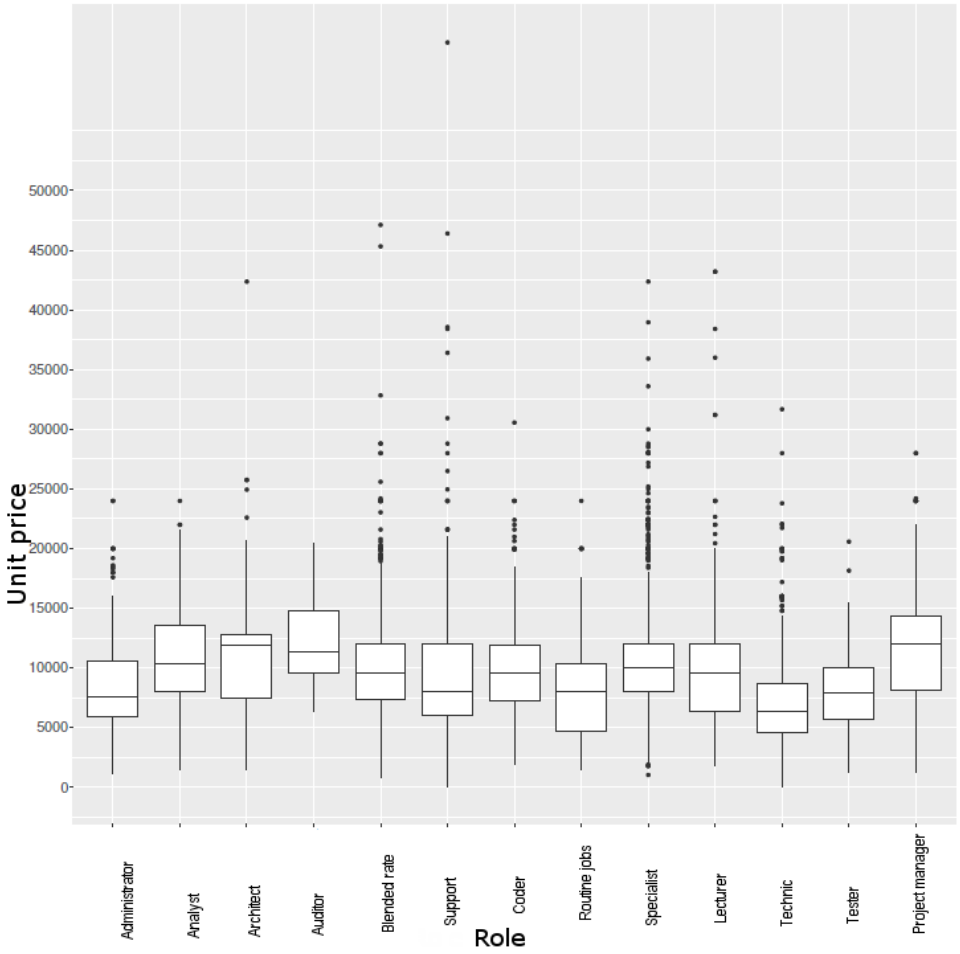


Figure 1. Summary of usual prices for roles.

The found usual prices are shown in Table 1. The graphical representation is depicted in Figure 1. As we can see, the result prices are slightly different for focal roles: The most expensive roles were project manager, architect and auditor, the least expensive was a technic. The role with the highest price variance was a support, which also brought the price with the absolute highest value.

A specialist was the role with the highest occurrence in the set. The area of specialization causes the difference in values within the role. An auditor (understood the IT or Information systems auditor) was the role with the least occurrence. The smallest price with eight Czech crowns per manday was stated in a contract with a high volume of other products and services, and hence probably the price was compensated in this case.

The detailed results, the dataset of all price data with links to original contracts, and the histograms and tables for selected roles are beyond this paper.

4. Discussion / Conclusion

While the numerical results of the research looked obvious and clear, there were aspects which should be considered when interpreting the results.

Majority of the accessed contracts were complex contracts which cover more than manday prices. Thus, the prices might be mutually compensated, which worsens objectivity of the result. Furthermore, the object of a contract is usually described in detail while the service description of the contract for the roles was often described very vaguely. On account of that and since the unit price was not cleansed from other pricing factors, it is not possible to make any conclusions about the real price for the selected role on the market.

Also, during the validation, the incorrectly extracted price data was corrected by the experts which was affected by a subjective view of the expert. The unclear cases could be (and was) classified by different experts differently.

Some contracts defined a unit manday price, although the roles were not named or further specified. For this reason, we defined a special category called Blended rate where the unit price is stated, but the role is indefinite.

Moreover, only contracts in the public sector are published in the registry, which was our only input the contracts between private subjects are confidential and not available for the research. We expected the prices in the public sector to be higher than in the private area. However, based on a discussion with the validation experts, the found unit prices tend to be rather lower than unit prices in private contracts while the overall price volume of the contracts tends to be higher. The reason could be the criteria of the selection, where the unit price is taken into account.

All those aspects decreased the possibility of generalization of the found price level to the whole market. The results only state which prices were contracted, and are very likely biased by the chosen details of the method. At least, the results could be useful for the statement of grounded theory for future research.

The risk is that the public sector workers may interpret the extracted usual price from the results as the price which is “desirable” and “normal” and use it not only for preliminary contract sizing for the tender but also for the negotiation with the vendors. This is accelerated by the government of the Czech Republic, which stated the list of usual prices as a mandatory consulting material for all state institutions public tenders.

References

Act No. 340/2015 Coll., on special conditions for the effectiveness of some contracts. Czech Republic.

Act. No. 536/1990 Coll., on prices. Czech Repiblic.

Brauer, F., Rieger, R., Mocan, A., & Barczynski, W. M. (2011). Enabling information extraction by inference of regular expressions from sample entities. In International Conference on Information and Knowledge Management, Proceedings (pp. 1285–1294). New York, New York, USA: ACM Press. https://doi.org/10.1145/2063576.2063763

Bruckner, T. (2019). Design of the technological architecture for PUMPIT project. Journal of Systems Integration 10(2) (2019) 34-40

Elastic Enterprise Search. (2020). Retrieved February 1, 2020, from https://www.elastic.co/enterprise-search

Mooney, R. J., & Bunescu, R. (2005). Mining knowledge from text using information extraction. ACM SIGKDD Explorations Newsletter, 7(1), 3–10. https://doi.org/10.1145/1089815.1089817

Ochrana, F., & Pavel, J. (2013). Analysis of the Impact of Transparency, Corruption, Openness in Competition and Tender Procedures on Public Procurement in the Czech Republic. Central European Journal of Public Policy, 7(2), 114–134.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Zobel, J., & Moffat, A. (2006). Inverted files for text search engines. ACM Computing Surveys, 38(2), 6-es. https://doi.org/10.1145/1132956.1132959

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11646

Communicating Corporate Social Responsibility through Twitter: a topic model analysis on selected companies

Camilla Salvatore[1], Annamaria Bianchi[2], Silvia Biffignandi[2]
[1]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy, [2]Department of Management, Economics and Quantitative Methods, University of Bergamo, Bergamo, Italy.

Abstract

Social media are fundamental in creating new opportunities for firms and they represent a relevant tool for the communication and the engagement with customers. The purpose of this paper is to analyse the communication of Corporate Social Responsibility (CSR) activities on Twitter. We consider the listed companies included in the Dow Jones Industrial Average Index and we implement a topic model analysis on their timelines. In order to identify the topic discussed, their correlation, and their evolution over time and sectors, we apply the Structural Topic Model algorithm, which allows estimating the model including document-level metadata. This model proves to be a powerful tool for topic detection and for estimating the effects of document-level metadata. Indeed, we find that the topics are overall well identified, and the model allows catching signals from the data. Finally, we discuss issues related to the validity of the analysis, including data quality problems.

Keywords: *Topic modelling; Structural Topic Model; Social media communication.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Social media are fundamental in creating new opportunities for firms and they represent a relevant tool for engaging with customers and stakeholder. Also the communication of corporate social responsibility (CSR) activities, which plays a fundamental role in enhancing firms' reputation, can enjoy the new opportunities deriving from their use (Cho, Furey, & Mohr, 2017). There is not a unique and shared definition of CSR. Table 1 provides an overview of the different classifications and shows that it is a multidimensional concept.

Table 1. CSR dimensions.

Reference	Dimensions
Carroll (1991)	Economic, Legal, Ethical and Philantropic
Dahlsrud (2008)	Environmental, Social, Economic, Stakeholder and Voluntariness
Kim et al. (2014)	Environmental, Philantropy, Education, Community/Employee involvement, public health, sponsorship of cultural/sports activities

Source: Amended from Cho et al. (2017).

Although computer-assisted analysis of CSR reports is common, the literature about the analysis of social media messages about CSR is scarce (Chae & Park, 2018). The purpose of this paper is to analyse the communication strategy of CSR activities through Twitter by a selected group of firms in order to answer the following questions.

Question 1. *Which CSR topics are discussed on Twitter?*

Question 2. *Which CSR topics are sector-specific?*

Question 3. *Which topics are likely to be discussed together?*

Question 4. *What is the topic evolution over time?*

The novelty of this paper lies in the following aspects. First, we focus on the messages posted by a selected group of companies rather than retrieving tweets that match a specific search query (group of relevant keywords). Second, for answering our questions, we apply the Structural Topic Model (STM) algorithm, which allows estimating the model including document-level metadata.

Section 2 introduces the model. In Section 3, the data and the model selection strategy are presented. The results are discussed in Section 4. The main conclusions are drawn in Section 5.

2. The Structural Topic Model (STM)

The STM is a probabilistic mixed membership model which allows to estimate a model including document-level metadata and, thus, to study the relationship between topics and metadata. In this section, we briefly describe the model; for further technical details, please refer to Roberts, Stewart, & Airoldi (2016), which originally proposed the model. This model is based on the *bag of words* representation, which means that each document is represented as a vector of words without giving importance to the order in which they appear. Let us consider a set of D documents indexed by $d \in \{1 \dots D\}$. Each document is composed by a mixture of words $w_{d,n}$, where $n \in \{1 \dots N_d\}$ indicates the position within the document. The collection of unique words is represented by a vocabulary. Each term in the vocabulary is indexed by $v \in \{1 \dots V\}$, it is assigned to a topic (z) and it is associated with the probability of belonging to each topic $k \in \{1 \dots K\}$. Thus, a topic is a mixture over words and the document is a mixture over topics. Document-metadata influence two components of the model, the *topical prevalence* that is defined as the proportion of the document that is associated to a topic, and *topical content* that refers to the usage rate of word in a topic. Thus, topical prevalence covariates affect the discussion proportion of the topic (θ), while topical content covariates affect the rate of word usage within a topic (β). The matrix of the P topic prevalence covariates and A topical content covariates are denoted by $X_{D\times P}$ and $Y_{D\times A}$ respectively. Model estimation and inference are based on a collapsed variational expectation-maximization algorithm. The model converges when the relative change in the approximate variational lower bound is below a defined tolerance level. Figure 1 summarizes the STM and highlights its three components: the *topic prevalence model* (left-hand side), the *topical content model* (right-hand side), and the *observation model* (central part).

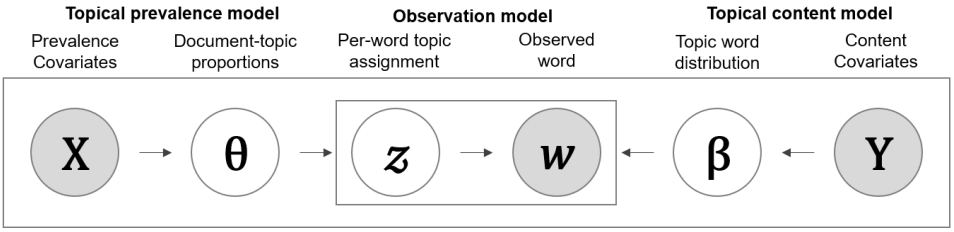


Figure 1. Structural Topic Model. Source: Amended from Roberts et al. (2016).

When estimating the model, the analyst must specify the algorithm initialization strategy and the number of topics. A shortcoming of topic models is that the output is very sensitive to the initialization. The *spectral initialization*, a deterministic algorithm based on the method of moments, is suggested due to its stability (Roberts, Stewart, & Tingley, 2019). Then, for choosing the optimal number of topics, it is necessary to compare some metrics. Roberts et al. (2019) argue that four metrics should be compared: residuals dispersion, held-out likelihood, semantic coherence and exclusivity. The held-out likelihood is a measure of

predictive power, which is useful for models comparison. The authors apply the document completion approach to estimate the held-out likelihood. The higher the held-out likelihood, the higher the model's predictive power. Taddy (2011) suggests that the dispersion of the residuals is one when the model is well specified. Residuals' dispersion is checked by means of a chi-squared test ($H_0: \sigma^2 = 1$ vs $H_1: \sigma^2 > 1$). A large number of topics should be preferred when rejecting the null. However, this requirement is very strict and for practical purposes, it is suggested to look at residual dispersion together with the other metrics. Mimno et al. (2011) present the concept of semantic coherence that is calculated for each topic k and it provides a measure of the co-appearance rate of the most probable words in that topic. If the most probable words in the topics tend to co-occur, then the topic is semantically coherent. Let $V^{(k)} = \left(v_1^{(k)}, \dots, v_M^{(k)}\right)$ be the list of the M most probable words in topic k. Then, define $D(v)$ as the *document frequency* for word v, and $D(v_m, v_l)$ as the *co-document frequency* for words v_m and v_l, i.e., the number of documents in which the selected terms occur together. Then, for each topic k, the semantic coherence is defined as follows:

$$C\left(k; V^{\{(k)\}}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^{(k)}, v_l^{(k)}\right) + 1}{D\left(v_l^{(k)}\right)}.$$

It is easy to check that semantic coherence will decrease as the number of topic increases, i.e., if the number of topics is small, it is likely that they will be composed by the same words. As countermeasure, Roberts et al. (2016) suggest to consider a measure of exclusivity, called FREX. Airoldi and Bischof (2016) develop this metric in a way that words frequency is balanced by exclusivity. Define $B(v^{(k)})$ as the *occurrence rate* of a word v in topic k. Then, for a set of comparison topics S, the exclusivity is defined as follows $E(k; v) = B(v^{(k)}) / \sum_{h \in S} B(v^{(h)})$. The FREX is defined for each topic k and term v as the weighted harmonic mean of term's frequency and exclusivity:

$$FREX_{k,v} = \left(\frac{w}{ECDF(B(v^{(k)}) / \sum_{h \in S} B(v^{(h)}))} + \frac{1 - w}{ECDF(B(v^{(k)}))} \right)^{-1}$$

where w is the weight in favour of exclusivity and ECDF stands for empirical cumulative distribution function.

3. Data and model selection

We selected the firms included in the Dow Jones Industrial Average index, i.e., a stock market index that measures the performance of the 30 largest US listed companies. We retrieved the full list of firms, joint with the activity sector from Bloomberg. Then, the original tweets (including retweets without a comment) posted on the firm's timeline have been collected. As reference period, we selected the second semester of 2019 (July-

December). Two firms (Apple and Walgreens) turned out not to have a Twitter account, while Walmart has been excluded due to rate limiting when retrieving data. The final sample includes 27 firms. Most of them operates in the Financial (18.5%), Technology (14.8%) and Health Care (14.8%) sectors. Then, there are Industrials and Consumer Discretionary (11.1% each) sectors, and Communication, Consumer Staples, Energy and Material ones (7.4% each). The number of messages retrieved is 8,602.

The stm R package developed by Roberts et al. (2019) has been used for implementing the analyses. The first step concerns the cleaning of the data. It involves different operations: elimination of punctuation, stop words, numbers, conversion to lower case, and stemming. The data are finally organized into documents, vocabulary terms and tokens (repeated words) as follows: 8,602 documents, 23,983 unique words and 136,201 tokens. After the cleaning process, only relevant terms remain. However, an additional step in data cleaning is the removal of infrequent terms (those that appear in a number of documents less or equal to a threshold). The threshold is defined as the number of documents in which the word appears. This operation is highly recommended because it allows reducing noise in the data, making the task of topic detection easier.

The choice of the appropriate threshold is made by looking at the number of the remaining documents, words and tokens (Table 2). Then, the analyst can assess the remaining terms in order to choose the appropriate threshold. For low values of the threshold, the reduction in the noise is small, thus we focused on higher values of thresholds, more specifically on 20, 30 and 50. After analysing the words that compose the vocabulary for each case, the most appropriate threshold seems to be 30.

Table 2. Comparison of thresholds.

Threshold	No. Documents	No. Words	No. Tokens
20	8591	1172	89663
30	8584	824	81024
50	8559	503	68486

Source: Authors' own elaboration.

The next steps are model specification and identification of the optimal number of topics. In our analysis, we only include topic prevalence covariates. We allow sectors and day to affect the discussion proportion of a topic. We estimate the day variable through a spline in order to account for non-linear effects. The optimal number of topics is chosen by looking at the metrics described in Section 2 (Figure 2, left-hand side). The appropriate number of topics seems to be around 40 and 50. It should be clear that there is no fixed way to choose among them, and this procedure does not yield the *true* number of topics. The differences in terms

of held-out likelihood and residuals dispersion are small. The trade-off between semantic coherence and exclusivity is evident. In order to choose among them, Figure 2 (right-hand side) compares the two metrics. The 32.5% of the 40 topics falls in the first quadrant, the 55% in the second one and the 12.5% in the fourth one. For the model with 50 topics, the percentages are 30%, 60%, 10% respectively. Thus, the model with 40 topics seems to be most appropriate.

4. Results and discussion

Topic discovery is performed by looking at the most-probable words for each topic, and labeling them consequently. We identify 21 topics related to CSR activities (Figure 3 left-hand side).

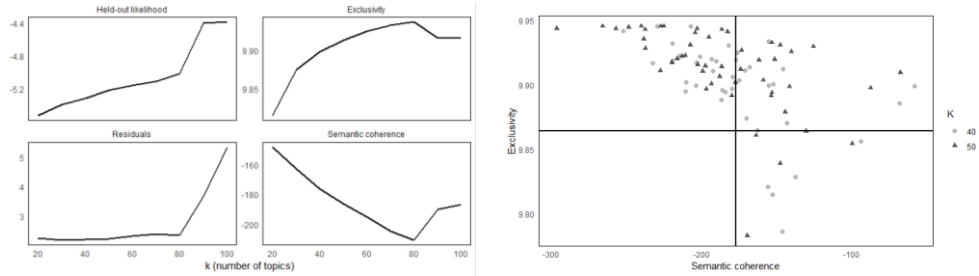


Figure 2. Evaluation metrics for choosing the number of topics (left) and comparison between exclusivity and semantic coherence for models with 40 and 50 topics (right). Source: Authors' own elaboration.

More precisely, 24.8% of them concerns the social dimension (community, employee engagement and sponsorship of events), 14.6% relates economic matters, 3.13% is on public health commitment, 2.13% concerns the environmental question, and finally, 1.95% of messages relates to educational programs.

The topics proportion of the identified CSR topics is plotted on the left-hand side of Figure 3. The topic correlation network is plotted on the right-hand side. It shows positively correlated topic, i.e., those topic that are likely to be discussed together within a tweet. Only correlations whose value is greater than 31% are plotted. Correlations within the same dimensions are evident. Moreover, two clusters have relevant features. Topics 39 (Education), 20 (Social), 25, 26 and 9 refers all to technological aspects. Topic 39 relates study programs involving technological instruments, Topic 20 relates the "digital transformation for helping communities" while the other topics are about the release of technological products or advertising about artificial intelligence, machine learning and digital services. The second cluster which includes "non CSR topics" (Topics 40, 19, 16, 33, 34, 1 and 17) relates advertising and promotions mainly linked to Christmas, Halloween and Summer holydays.

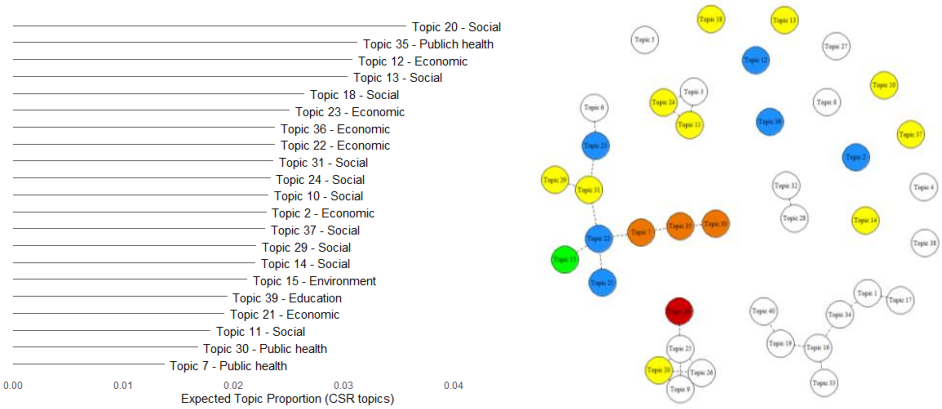


Figure 3. Expected topic proportions of CSR topics (left) and topics correlation (right) with sector indication (Green for Environment, Red for Education, Yellow for Social, Blue for Economics, Orange for Public health).

The novelty introduced by this model is the possibility of estimating the effects of topical prevalence covariates on the discussion proportion of a topic. We start from the sector variable (Figure 4). It is not surprising that firms in the energy and materials sectors tweet significantly more about environmental issues than the others (Topic 15). Interesting patterns can be observed for Topic 37 that concerns events sponsorship, mainly of the *#voteyourmainstreet* initiative. This event was sponsored by American Express, which belongs to the financial sector, i.e., the one that tweeted significantly the most. The second topical prevalence covariate is time. Figure 5 shows the expected topic proportion as a smooth function of the day with 95% confidence intervals. Topic 15 remained stable over time, with a small reduction during summer and winter holidays. Topic 37 shows a higher proportion during October and November, the months when the sponsored event mainly took place. Finally, Topic 10 that concerns supporting small businesses has a peak in the last days of November and the beginning of December. Indeed, in that period the *Small Business Saturday* initiative took place, that is a traditional event to support small businesses and for celebrating communities.

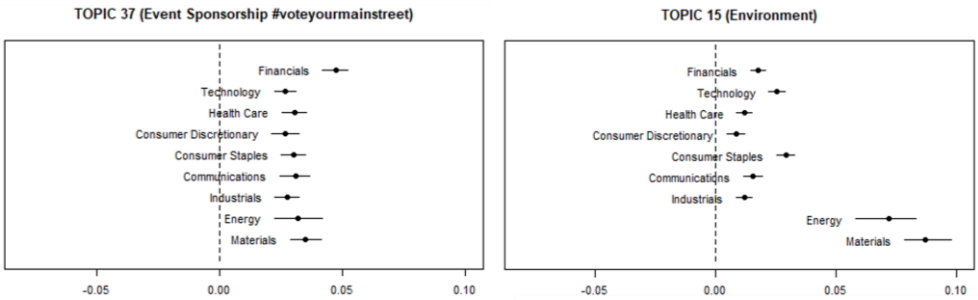


Figure 4. Effect of the "sector" on the proportion of topic discussion. Source: Authors' own elaboration.

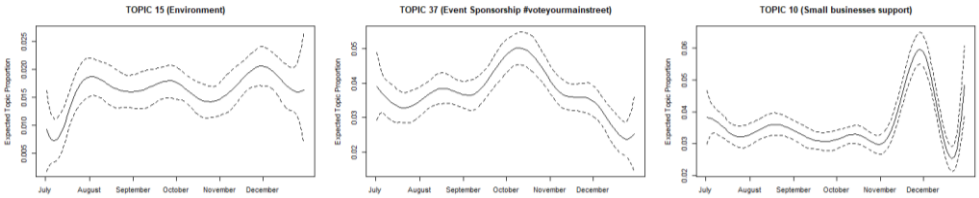


Figure 5. Expected topic proportion over time. Source: Authors' own elaboration.

5. Conclusions

In this work, we propose to apply the STM model for analyzing the communication of a selected group of firms about CSR activities on Twitter, allowing topical prevalence to evolve over time and varying across sectors. With reference to the initial questions, STM proves to be a powerful tool for topic detection and for estimating the effects of document-level metadata. Indeed, we get evidence that some topics are sector-specific and that the model allows to catch signals from the data, in correspondence of particular events. In addition, interesting correlations have been highlighted. When analyzing Twitter data, practitioners should be aware about data quality aspects and the errors they may encounter (Salvatore, Biffignandi, & Bianchi, 2020). Indeed, the main shortcoming is that the output of the analysis is very sensitive to the analyst' judgements at the various steps. Further developments may concern the analysis of data quality aspects, the inclusion of covariates' interactions, and of topical content metadata.

References

Airoldi, E., & Bischof, J. (2016). Improving and Evaluating Topic Models and Other Models of Text. *Journal of the American Statistical Association, 111*, 1381-1403.

Carroll, A. (1991, 7 1). The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business Horizons, 34*(4), 39-48.

Chae, B., & Park, E. (2018, 6 28). Corporate Social Responsibility (CSR): A Survey of Topics and Trends Using Twitter Data and Topic Modeling. *Sustainability, 10*(7), 2231.

Cho, M., Furey, L., & Mohr, T. (2017). Communicating corporate social responsibility on social media: Strategies, stakeholders, and public engagement on corporate facebook. *Business and Professional Communication Quarterly, 80*(1), 52-69.

Dahlsrud, A. (2008, 1 1). How corporate social responsibility is defined: an analysis of 37 definitions. *Corporate Social Responsibility and Environmental Management, 15*(1), 1-13.

Kim, S., Kim, S., & Sung, K. (2014, 10 28). Fortune 100 companies' Facebook strategies: Corporate ability versus social responsibility. *Journal of Communication Management, 18*(4), 343-362.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & Mccallum, A. (2011). *Optimizing Semantic Coherence in Topic Models.* Association for Computational Linguistics.

Roberts, M., Stewart, B., & Airoldi, E. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association, 111*, 988-1003.

Roberts, M., Stewart, B., & Tingley, D. (2019, 10 31). Stm: An R package for structural topic models. *Journal of Statistical Software, 91*(1), 1-40.

Salvatore, C., Biffignandi, S., & Bianchi, A. (2020). Social Media and Twitter Data Quality for New Social Indicators. *Social Indicators Research*. doi:10.1007/s11205-020-02296-w

Taddy, M. (2011, 9 21). On Estimation and Selection for Topic Models. *Journal of Machine Learning Research, 22*, 1184-1193.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11647

Proposal of a composite indicator for measuring social media presence in the wine market

Andrea Conchado Peiró[1], José Miguel Carot Sierra[1], Elena Vázquez Barrachina[1], Enrique Orduña-Malea[2]
[1]Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Spain, [2]Department of Audiovisual Communication, Documentation and History of Art, Universitat Politècnica de València, Spain.

Abstract
Cybermetrics field is attracting considerable interest due to its utility as a data-oriented technique for research, though it may provide misleading information when used in complex systems. This paper outlines a new approach to market research analysis through the definition of composite indicators for cybermetrics, applied to the Spanish wine market. Our findings show that the majority of cellars were present in only one or two social media networks: Facebook, Twitter or both. Besides, the presence on the Web can be summarized into three principal components: website quality, presence on Facebook, and presence on Twitter. Three groups of cellars were identified according to their position in these components: cellars with a high number of errors in their website with complete absence of information in social media, cellars with strong presence in social media, and cellars in an intermediate position. Our results constitute an excellent initial step towards the definition of a methodology for building composite indicators in cybermetrics. From a practical approach, these indicators may encourage cellar managers to make better decisions towards their transition to the digital market.

Keywords: *Composite indicator; principal component analysis; wine; cybermetrics; social media.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

There is evidence of a growing interest in research related to the impact of new products, brands and firms through their websites and social media profiles (Orduna-Malea and Alonso-Arroyo, 2017). Most of these studies have been traditionally based on link analysis or Search Engine Optimization (SEO) techniques, usually with a marketing-oriented approach. However, the majority of these analyses do not follow empirically tested or validated methodologies by the scientific community. On the contrary, they are mainly focused on narrow preestablished objectives or specific case studies, such as the analysis of a unique brand, biased studies skipping other potential emerging competitors or qualitative analysis with non-representative customer profiles.

Cybermetrics is the field that studies how to build and use web resources, structures and technologies (Björneborn e Ingwersen, 2004). Quantitative advances in this area are mostly addressed to solve research problems in social sciences through the application of quantitative research methods (Thelwall, 2009). The huge dependence on the availabily and variability of both metrics and sources has jeopardized the use of Cybermetrics (Thelwall, 2010), mostly limited to link analysis. However, this particular technique is insufficient when it comes to analyse complex environments, which cannot be describe with simple metrics.

Recent developments have led to the design and assessment of measurement systems composed by different indicators, with the aim of comparing results within different units of analysis (Orduna-Malea and Alonso-Arroyo, 2017). These measurement tools are based on composite indicators (or indexes), which are defined as the combination or mathematical aggregation of a set of simple indicators aiming to summarize a multidimensional concept into a simple or one – dimensional index, based on an underlying theoretical model (Nardo et al., 2008). This system of composite indicators represents a sound procedure for measuring websites' impact on the basis of its ability to identify profiles and trends as regards to the supply (contents generated by companies) and demand (contents searched by users) in particular economic sectors. This method might help to provide useful information for the commercialization of new products. This work precisely aims to apply this method to the Spanish wine sector.

Recent evidences show that the international wine market can be divided into two clusters: countries with a traditional approach as regards wine commercialization, like France, Italy and Spain (Old World), and countries with stronger presence in online international markets, such as United States, Argentina, Chile, South Africa and New Zeland (New World). Following that logic, the presence of Spanish cellars both on the Web and social media channels is expected to be quite low, as previous findings have evidenced (Compés and Castillo, 2014).

In the light of these events, the main objective of this paper is to outline a new approach to market research analysis through the definition of Cybermetric composite indicators, to be applied to the Spanish wine market.

This work is based on the experience of the authors as the coordinator team of the research funded project: 'eMarketwine – Design of a method and an online tool for information intelligence addressed to geolocalized recommendations in the wine industry' (Ref. CSO2016-78775-R). The Spanish wine sector (including only bottled wine) was chosen because it is one of the most relevant sectors of the Spanish industry. This project follows on a previous project, 'Trademetrics – Methodological proposal of a cybermetric analysis of products, branches, people and firms in the Spanish online market' (Ref. CSO2013-46138-P), also funded by the Ministry of Economy and Competitiveness.

2. Methodology

2.1. Data

The sample consisted of 3,164 Spanish cellars, collected between 2018 and 2019. As a result, 23 variables were collected, of which seven were quantitative variables (metrics), and the rest were qualitative variables (Yes / No).

Metrics gathered are divided into two main categories. First, metrics related to the technical quality of cellars' websites. These data were obtained from W3C Markup Validation Service and Link Checker (https://validator.w3.org). Second, metrics related to the activity of cellars on Twitter and Facebook. These data was extracted via API. The type, description and nature of all the variables considered for each cellar are included in Table 1.

Table 1. Type, label, description and nature of variables analysed.

Type	Variable	Description	Nature
Identification	Id	Numeric code	Numeric
	Name	Name of cellar	Character
Characteristics	Zip code	Location	Character
	D.O.	Denomination of origin	Qualitative
Metrics	broken_links	Number of broken links	Quantitative (integer)
	html_errors	Number of HTML error codes	
	html_warnings	Number of HTML warnings	
	Likes	Number of Likes in Facebook	
	fb_followers	Number of Facebook followers	
	tw_posts	Number of posts in Twitter	
	tw_followers	Number of Twitter followers	
Social media presence	Facebook	The cellar has a profile on …	Qualitative (Yes/No)
	Twitter		
	Linkedin		
	Pinterest		
	Flickr		
	Youtube		
	Instagram		

2.2. Methods for building composite indicators

Principal Component Analysis (PCA) was chosen to analyse the seven quantitative variables, since it is one of the most practical ways to measure constructs that cannot be directly measured (so–called latent variables). This technique is useful for understanding the structure of a set of variables, such as the set quantitative metrics considered. Besides, it provides a tool for reducing this data set to a more manageable size, while retaining as much of the original information as possible.

This data analysis technique is concerned with establishing the underlying linear components that exist within the data, and how much each variable contributes to each component. With this aim, the eigenvalues of the correlation matrix representating the relationship between

variables are calculated. These eigenvalues are subsequently used to calculate eigenvectors in such a way that eigenvalues represente a measure of the substantive importance of the associated eigenvector. According to Kaiser (1960), only components with eigenvalues greater than 1 should be retained. This simple rule of thumb has generated a high number of recommendations about the appropriate number of components to retain. Once factors have been extracted, factor rotation procedures rotate factor axes such that variables are loaded maximally on only one factor. Among available rotation procedures, we can chose between orthogonal methods (including varimax, equamax or quartimax procedures) or oblique rotation, depending on whether components are allowed to correlate.

Log–transformation was needed for dealing with the highly skewed data contained in metrics. Next, normalization procedures were applied in order to guarantee the robustness of the analysis against the presence of outliers in the data (Ebert & Welsch, 2004). Thanks to this transformation, the composite indicator was comparable across different groups. Among all available methods, min–max normalization was chosen due to the simplicity of interpretation of the resulting composite indicator.

3. Results and discussion

3.1. Quality of data in metrics (technical website quality and social media activity)

As expected, Facebook and Twitter were the social media networks with a stronger presence of Spanish cellars. Figure 1 shows that other networks had neglible presence of cellars compared to this results, like Flickr, Pinterest or Linkedin. The majority of cellars were present in only one or two social media networks.

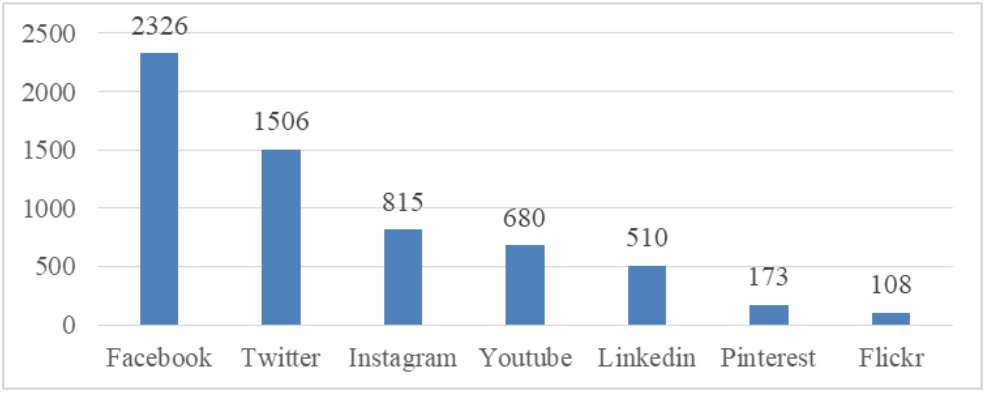


Figure 1. Number of cellars present in each social networking site.

There was a high level of heterogeneity among data metrics concerning technical quality of the website, intensity or volume of use in Facebook and Twitter (Table 2). There was a high

level of correlation between the number of Facebook's likes and Facebook's followers (r ~ 1), and a moderate level of correlation between the number of Twitter's posts and Facebook's followers (r = 0.67). Due to the high number of outliers with a high number of followers both in Facebook and Twitter, data metrics presented high level of asymmetry and kurtosis.

Table 2. Quality of data metrics.

Parameter	**Website**			**Social media**			
		Technical quality		**Facebook**		**Twitter**	
	Broken links	**HTML error**	**HTML warnings**	**Likes**	**Followers**	**Likes**	**Followers**
Valid (%)	90%	93%	93%	60%	60%	46%	46%
Min	0	0	0	0	0	0	0
Max	197	790	377	381,000	381,000	38,717	24,017
Mean	8	24	24	2,632	2,677	1,260	1,337
SD	13	54	28	14,249	14,266	2,564	2,480
Asim Std.	5	6	3	20	20	7	4
Kurt Std.	49	57	22	458	456	66	24

Among the data set composed of 3,164 cellars, we selected those cellars with valid information in data metrics (N = 3,052) and presence in social media networks (N = 2,478). Next, we applied PCA to the data set of seven metrics, with varimax rotation procedures and pairwise selection methods for missing data (KMO = 0.57, Barlett's test $\boldsymbol{X^2 = 12.487, p = 0.000}$). As a result, three principal components were extracted which explained 72.5% of variability contained within the initial data set.

Table 2. Quality of data metrics.

Metrics	PC1	PC2	PC3
broken_links	0.006	0.168	0.427
html_errors	-0.005	-0.035	0.730
html_warnings	0.003	-0.044	0.789
fb_likes	0.987	0.157	0.001
fb_followers	0.987	0.159	0.001
tw_ followers	0.130	0.907	-0.009
tw_post	0.175	0.881	0.117

According to the loading scores obtained through PCA, the first principal component (PC1) was labelled as 'Presence in Facebook', PC2 was labelled as 'Presence in Twitter' and PC3 was labelled as 'Website's technical quality'. Consistently with the correlation coefficient, both Likes in Facebook and Followers in Facebook showed similar values for their loading scores, whereas Followers in Twitter presented a higher contribution in its corresponding principal component. As regards to the third component, websites' technical quality, the number of broken links presented the lowest contribution to these latent variables, as evidenced in the low value of communality (0.189). All other values of communalities were higher than 0.5.

Thus, the findings using PCA confirm the strength of the association between similar variables concerning these three principal components. According to these results, it can be concluded that each principal component can be summarized through its scores. However, observed variables are preferred for building composite indicators, rather than latent variables, as those obtained with PCA. Thus, accordingly we decided to select the number of followers in Facebook and Twitter to build the composite indicator. Both variables presented high values of communalities resulting from the extraction of the previous three components. Besides, both were referred to people using different social media which resulted in higher values of internal consistency (4 items, Cronbach's $\alpha = 0.714$) than the solution including items concerning websites' technical quality (7 items, Cronbach's $\alpha = 0.649$), despite the low number of items.

As previously shown, variables about presence in Facebook and Twietter were highly skewed due to the presence of outliers. Thus, log–transformation of data was applied in order to deal with these non–normal distributions. Additionally, min-max normalization was also used for building a composite indicator with two components, each of them weighted using a rank between 0 and 50. This latter decision was taken as a practical way to generate an indicator

with a range of variation of 100 points, through the sum of two different components with independent ranges from 0 to 50. This method of weighting was chosen because it was the most practical ways to create a meaningful indicator to be used and applied by managers and steering committees of commercial wine cellars. Though many other methods of weigthing are available for composite indicators, researchers have not reached an agreement about the criteria to priorize among them. Thus, we considered different solutions and finally decided to follow this procedure on account of its feasibility for this environment.

Thus, the resulting formula for our proposal of a composite indicator measuring presence in social media networks is the following:

$$\boldsymbol{IFT_i = 50\frac{LogF_{fi}}{\max_i(LogF_{fi})} + 50\frac{LogT_{fi}}{\max_i(LogT_{fi})}}$$

Where:

F_f= Followers on Facebook

T_f= Followers on Twitter

The graphical representation of the IFT_i score of all cellars in this composite indicator in a scatterplot with websites' technical quality allows the reader to identify three groups of cellars (Figure 2). First, cellars with a high number of errors in their website with complete absence of information in social media. These non-digital companies have been marked in red. Second, and on the opposite side, there is another group of cellars with a strong presence in social media. The majority of them also have high quality websites (in terms of technical developement), though some outliers might consider to invest in the technical improvement of their websites. Third, in an intermediate position, we find cellars with moderate presence in social media and different levels of technical quality in their websites.

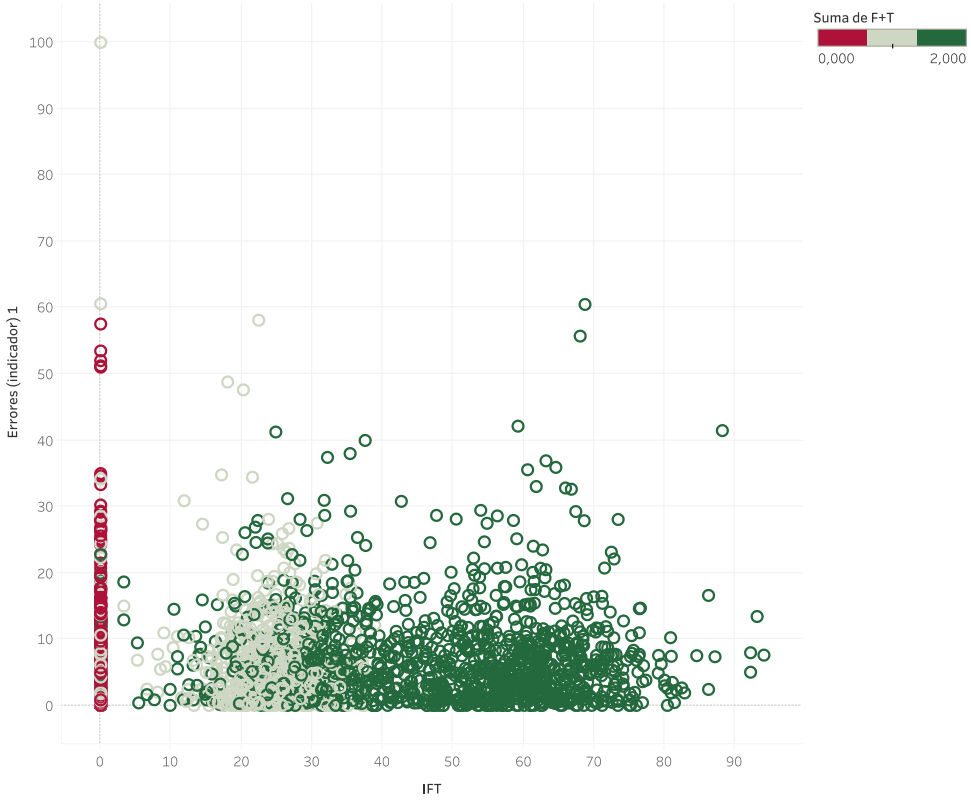


Figure 2. Scatterplot of IFT versus Websites' technical quality.

4. Conclusions

This study has highlighted the importance of composite indicators in cybermetrics when dealing with highly skewed data and the need to apply specific methods of data analysis for extracting useful information in decision making. We have devised a methodology which can be useful in other research areas concerning cybermetrics. This work has some limitations, such as the reduced number of metrics involved in the analysis. Research is underway to overcome this shortcoming. However, in our view, these results constitute an excellent initial step towards the definition of a methodology for building composite indicators in cybermetrics.

Specifically, results have evidenced that activity in social networking sites is not strongly related to the websites' technical quality for Spanish cellars. This result is unexpected as it could be assumed a priori that top companies would be good elsewhere, either creating websites without errors or showing high activity in the social media channels. Otherwise, the presence of Spanish cellars has been shown to be concentrated in Facebook and Twitter.

Moreover, the number of followers in these two social networking sites is enough to explain Spanish cellars' impact variability on these online spaces.

These results have important managenerial implications for cellars' managers, who can monitor the market activity on social media, and systematically carry out benchmarking analyses using robust statistical methods. Likewise, results can be used for advisory and consultancy activities oriented to strategic decision making and online marketing. The future inclusion of aggregated data, such as denomination of origin or region, will expand the capabilities of the system.

Acknowledgements

This work was carried out within the framework of a Spanish research project 'eMarketwine: diseño de un método y una herramienta de online information intelligence orientada a la recomendación geolocalizada para el mercado del vino' (Ref. CSO2016-78775-R), founded by the Ministerio de Economía y Competitivad (Spanish Ministry of Economy and Competitiveness).

References

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Jasist*, 55(14), 1216-1227.

Compés López, R., & Castillo Valero, J.S. (2014). *La economía del vino en España y en el mundo*. Alicante: Cajamar Caja Rural.

Ebert, U., & Welsch, H. (2004). Meaningful environmental indices: a social choice approach. *Journal of Environmental Economics and Management*, 47(2), 270-283.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, 26(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1), 141-151.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2008). *Handbook on constructing composite indicators: methodology and user guide*, OECD Statistics Working Paper, STD/DOC (2005)3, OECD Publishing, Paris.

Orduna-Malea, E., & Alonso-Arroyo, A. (2017). *Cybermetric techniques to evaluate organizations using web-based data*. Cambridge: Elsevier.

Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences*. San Rafael, CA: Morgan Claypool.

Thelwall, M. (2010). Webometrics: Emergent or Doomed?. *Information Research*, 15(4), 1-10.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11650

Political Polarization and Movie Ratings: Web Scraping The Brazilian Contemporary Scenario

Ana Paula Moritz[1], Bruno Chagas[2]
[1]Department of Social Sciences, Pontifical Catholic University of Rio de Janeiro, Brazil,
[2]Department of Computer Science, Federal University of Minas Gerais, Brazil.

Abstract

Our work aims to analyze the impact of political polarization on movie ratings at the IMDb platform. For that we explore the concepts of Word of Mouth and Buzz marking the important role they play on polarized opinions in movie ratings.We develop a code on python to perform web-scrapping on the sample scope of Brazilian movies and interpret the data collected using a controversiality index based on standard deviation. The outcome sheds some light into the relation between Buzz and Controversiality within the framework of Brazil's current political scenario.

Keywords: *Political Polarization; Movie Ratings; Web-Scrapping; Brazil; Controversiality; Buzz.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

The idea for this paper came with the urge to explore data related to polarization in public opinion, specifically when it comes to cultural products. In order to do so, we have chosen to web-scrap the International Movie Database (IMDb) platform in search for the most controversial movies in Brazil. Our idea is to point out how political opinions are associated with a controversiality index.

Some papers in this area have already been written, but none of them refer to the political aspects of polarized opinions in movie ratings. In Fischoff, Antonio, Lewis (1998), the authors explore favorite films and film genres as a function of race, age and gender; in Wühr, Lange, Schwarz (2017) the authors compare gender stereotypes with actual movie preferences; Oliver et al. (1998) observe the impact of sex and gender role self-perception in reactions to different types of movies; Koh, Hu, Clemons (2010) investigate movie reviews online and mark the cultural aspects of quality perception; Otterbacher (2012) analyzes writing styles and metadata features on movie review forums focusing on gendered comments. All referred papers examine a variety of identity markers to explain partiality in movie preferences, but the political aspects of polarization is something we believe should also be taken into consideration. It's still unclear to us whether this phenomenon happens only in Brazil, or if it was overlooked by previous research.

In Brazil, polarization has been a big part of society's everyday life since 2013, approximately. The country went through a powerful but somewhat diffuse street movement that claimed to be apolitical. The outcomes of this movement prove otherwise. There was a fracture in society that had just been exposed and, since then, the country was never the same. After an impeachment in 2016 -considered by some as a coup d'état- and electing a far-right, ultra-conservative president in 2018. Brazil currently has its most conservative and religion-oriented government. In case readers want a deeper analysis on the Brazilian political situation, we suggest Chagas-Bastos (2019) and Pinheiro-Machado (2014).

We observed the controversy generated by some movies in social media, in order to place and compare them with the data we've collected. Our methodological approach to this issue was to develop a code for web-scraping the IMDb page, we searched for movies with the highest numbers of votes, then analyzed the normalized standard deviation, correlated the results with our hypotheis and conceptualized it with applicable literature. In section 2 we describe the concepts of eWoM and Buzz and their association with polarized opinions as well as deepen our explanation of the web scraping process and controversiality index. In section 3 we present our results and analyze them further, connecting data, theory and scenario. In the last section we present our conclusions and remarks and also make suggestions on how to broaden this research.

2. Theoretical Background

In order to explain our hypothesis, we explored the concepts of Word of Mouth (WoM) and Buzz, as well as their influence on polarized opinions and movie ratings. In the first part of this section we demonstrate how they work and are instrumented in order to mobilize voters - by doing so we give life and meaning to the statistical facts. As for the second part of this section, we present our method for constructing the code on python, and how we applied it in our analysis. We also present the concept of hard controversiality that verses on love-hate movies and how it frames polarization in movie ratings.

2.1. eWoM & Buzz

The concept of WoM is defined by Richins (1984) and Sundaram (1998) as a way people influence each other through communication. Liu (2006) and Mohr (2007) apply this concept to the movie industry addressing box office revenue and marketing strategies for movie premieres. We apply the concept as an informal influence, exerted by opinion leaders that is based on interpersonal relations in an online environment, it is named eWoM.

As stated by Moon (2010), eWoM can act positively or negatively, depending on the movie and its public, the author enumerates five factors that should be observed when evaluating movie ratings and recommendations such as (1) number of ratings, (2) average rating, (3) rating standard deviation, (4) percentage of highest rating, (5) percentage of lowest rating. First, the author explains that the accumulated number of ratings would indicate how many people have seen the movie, as we will demonstrate further in our analysis, this behavior doesn't occur in the data we have collected. Second, the average rating means how viewers rate good movies high and bad movies low, but we can also analyze ratings by a evaluation of disagreement towards the same movie given by the standard deviation.

Hennig-Thurau et al. (2004) categorized what motivates people to articulate their opinion online, the main features of eWoM presented in this paper are: focus-related utility, consumption utility and approval utility. The first, and most important to our hypothesis is related to adding value to community, it could be interpreted as helping a political party, a politician or an ideology. The ones that are motivated by this, expect social and economic benefits out of their actions. The second consists of using the websites to give opinions and gather information about products, but it does not apply to our case, since IMDb doesn`t carry this interactive feature for its users. The third one is about self-enhancement and economic reward; users expect to be recognized and praised for their opinions and in some cases, users receive compensations as tokens of appreciation by the reward giver. Moreover, the authors also indicate some discursive practices such as vengeance, boycott, altruism, exertion of power and frustration.

The definition of Buzz in Mohr (2007) consists on "the practice of gathering volunteers formally either by actively recruiting individuals who naturally set cultural trends, or informally by drawing 'connectors': people who have lots of contacts in different circles, who can talk up their experiences with folks they meet in their daily lives". Some of the profiles the author refers to are: experts, members of the press, politicians, celebrities, or well-connected customers others rely on for information. In the Brazilian scenario we can observe different cultural patterns such as evangelical leaders influencing entire congregations to give bad reviews on a movie they believe is threatening to their faith or political militants who are often hired to spread Buzz and count with the help of bots on social media platforms in order to reach more people faster.

We noticed that previous literature doesn't cover the political component present on eWoM and Buzz, we argue that these concepts per se do not spawn polarization, there is an exterior factor that when injected into the debate sparks a fire. Kostakos (2009) evidences a strong bias when it comes to the crowd's wisdom, and also in the design of social media platforms, as demonstrated by Balasubramanian and Mahajan (2001) in virtual communities. An important feature of the Brazilian scenario is the extent of social media usage for debates and comments, and how strongly polarized opinions dispute control over a narrative that might influence the public's opinion.

2.2. Web-Scrap and Controversiality

We constructed a code on python language that uses the proper packages for web-scrap. Our purpose is to evaluate and analyze some relevant statistical components over the IMDB platform, which allowed us to study controversiality and engagement properties. We considered movies with at least 900 ratings, only brazilian productions, spoken in portuguese. Our sample performed a total of 214 feature movies with around 1,8 million individual ratings. The data was collected by the end of February 2020.

IMDB platform displays just the mean value of a specific movie. Although, as seen in Hoßfeld, Schatz and Egger (2011), solely the mean value causes a "loss of important information on the user rating diversity". In order to retain more information, we used a statistical component called normalized standard deviation or hard controversiality index proposed by Amendola, Marra and Quartin (2015) which allowed us to seek love-hate movies. We will call it by C throughout our text. Therefore, we can write the controversiality index C in the form:

$$C = \frac{1}{C_{max}} \left[\sum_{i=1}^{10} p_i (r_i - \overline{r})^2 \right]^{1/2}$$

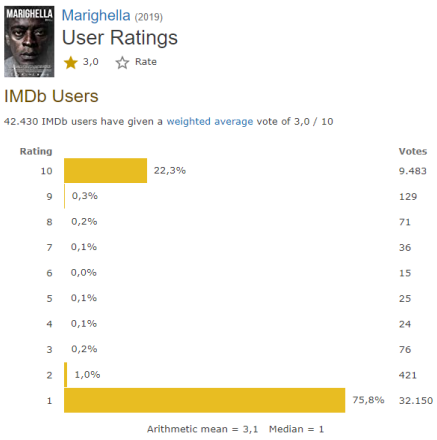
where the index i varies over the number of ratings, r_i is the rating i, p_i the percentage of ratings r_i and $\overline{r}$ is the mean value and Cmax is the maximum value that C can be achieved. The maximum number is 4,5, and we can see this effect when the polarization is maximum: 50% of ratings in 1 and 50% in 10 and this is the maximum love-hate for a movie.

For our next section we will analyze some aspects of C and its relation to the number of individual ratings on a movie. We argue that there is an engagement factor, in which high values of C and ratings indicate some kind of mobilization. Furthermore, they shed some light on how WoM and Buzz are correlated to C and the amount of ratings.

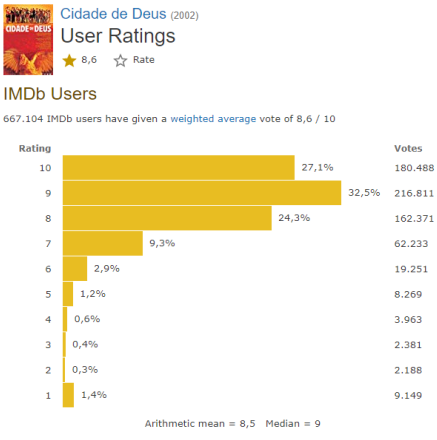
3. Results and Analysis

We start our analysis illustrating the phenomenon of controversiality and polarization on movie ratings with the help of figure 1(a) and figure 1(b). As we can see, the first one depicts a highly controversial movie with C≅0.84, while the second has a normal (or expected) behavior with C≅0.35. We stand that all movies with a political component, internal or external, in the current Brazilian political scenario, are subjected to biased public scrutiny and we may say that this is a Buzz-oriented phenomenon. It is important to point out that this pattern repeats itself .

We highlight that figure 1(a) is our key example, and its characteristics permeate the other movies we have analyzed. A peculiar trait of this movie is that its premiere in 2019 was censored in Brazil and it is expected to be released in May 2020. Here we outline the main question of our paper: how come a movie that wasn't even released had already the 5th largest number of votes and the 4th highest C? Therefore, we assert that the political component in movie ratings, as described by us, is often overlooked throughout the existing literature.



(a)Marighella (2019)



(b) Cidade de Deus (2002)

Figure 1.Statistical data from two brazilian movies. Source:IMDb (2020).

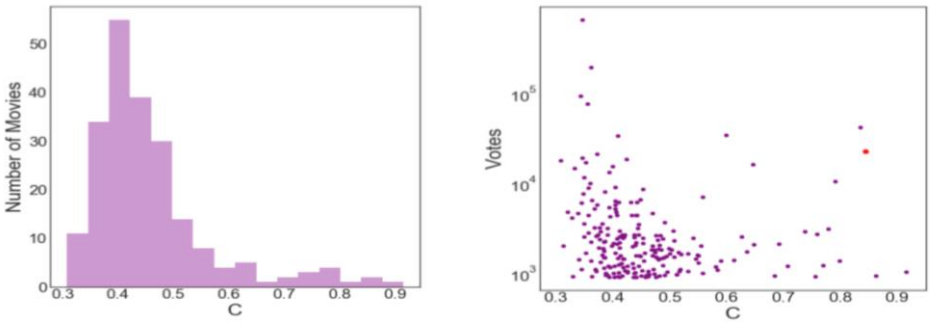


Figure 2. Number of movies and votes as function of C on our web-scrap.

Figure 2 on the left depicts that C the number of movies do not converge to zero while C is increasing as normal distribution. Moreover, we can notice a bump around 0,8 and 0,6 along the C axis, pointing out some external cause to it. Figure 2 on the right shows us that there are some movies that have a higher C and number of votes, we will analyze as organized in table 1.

Table 1. Most controversial movies with higher number os votes.

Movie (Year)	***Votes***	***C***	***Movie (Year)***	***Votes***	***C***
Marighella (2019)	*42.395*	*0,84*	*The First Temptation of Christ (2019)*	*2.755*	*0,76*
Nothing to Lose (2018)	*22.777*	*0,84*	*Lula, O Filho do Brasil (2009)*	*2.140*	*0,70*
Aquarius (2016)	*16.412*	*0,65*	*Nosso Lar (2010)*	*2.568*	*0.63*
The Edge of Democracy (2019)	*10.582*	*0,80*	*Marighella (2012)*	*2.954*	*0,74*
O Mecanismo (2018)	*34.753*	*0,60*	*Polícia Federal: A Lei é para Todos (2017)*	*3.154*	*0,70*

We have organized Table 1 according to the highest values of C and the highest number of votes in order to evidence where the cluster of polarization occurs. It is noticeable that the movies that generate the highest C and mobilize a larger number of votes are quite recent. Our hypothesis is that the political turmoil that Brazil went through in 2013 led the country into a generalized social and political instability which affected people's perspectives regarding cultural products in general, and more specifically as we show, movies. We sustain our hypothesis with Aquarius, the movie was nominated for a Palme d'Or in Cannes, and at the award ceremony the cast protested against the 2016 coup d'état - this action is what sparked the buzz - then the movie and cast were fiercely criticized which granted the movie its controversial status.

Returning to our analysis on Hennig-Thurau et al. (2004), we would like to recodify and broaden the utility types for engaging. When the authors categorize a motive as "adding value to community" and the person is motivated by the idea of helping the company, we can interpret that as a portion of society that wants to help their country, ideology or political party, in short, their motivation is partisanship. People who identify with the right-wing, envision the country as a corrupted institution and want to save it by restoring its traditional moral values. People who identify with the left-wing are trying to restore what the country once was during the leftist governments. To that end, what Hennig-Thurau et al. (2004) call altruism, here we can describe as an ideological component.

This ideological component is the answer to our analysis of the correlation between eWoM and Buzz with C. The movies listed below are in some way associated with either the right-wing (O Mecanismo, Polícia Federal: A Lei é para Todos and Nothing to Lose) or the left-

wing (The Edge of Democracy, Lula, O Filho do Brasil, Marighella, Aquarius and The First Temptation of Christ). The only outsider is Nosso Lar, a movie based on spiritism, a minoritarian religion in Brazil that suffers with persecution and intolerance.

4. Conclusion and Remarks

In summary, we have presented evidence on the correlation between the concepts of eWoM and Buzz with the controversiality index. The proportion and motivation of public engagement is boosted by the influence of the Buzz in social media platforms. Our results indicate that the political and ideological components affect and bias public scrutiny when it comes to movies and ratings. We believe that this paper is relevant for it opens possibilities for deeper analysis of a political and ideological component in movie preferences, something that hasn't been done before. As in present time political disputes tend to become more vicious and intense, we insist on the observance of democratic principles, and affirm the importance of caution with eWoM and Buzz spreaders.

Results so far have been encouraging, we intend on expanding this research in order to fill some gaps, and apply this method to other cultural products and countries where political polarization is on the rise. We also consider expandidng the research to other platforms, following Kostakos (2009) hypothesis that the design of platforms favours biased opinions. We would also like to explore the action of bots and botnets, as it was an element that we weren't expecting on encountering and it prompted some consistent questions as to: who orchestrated their actions and why, if there was any financial gains involved and what's the purpose behind that. Another perspective for this analysis that didn't fit our research scope is text-mining via sentiment analysis and summarization techniques, through that we expect to find the most important or commonly used words, and if they fit a political and ideological context or agenda.

References

Amendola, L., Marra, V., & Quartin, M. (2015). The evolving perception of controversial movies. *Palgrave Commun, 1*.

Balasubramanian, S., Mahajan, V. (2001). The economic leverage of the virtual community. *International Journal of Electronic Commerce, 5(3).*

Chagas-Bastos, F. (2019). Political realignment in Brazil: Jair Bolsonaro and the right turn. Revista de Estudos Sociales [Online], 69.

Dent, A., Pinheiro-Machado, R. (2014). Introduction: the cellularity and continuity of protest in Brazil. Anthropological Quaterly, Vol. 84, No. 3, pp. 883-885.

Fischoff, S., Antonio, J., Lewis, D. (1998). Favorite films genres as a function of race, age, and gender. *Journal of Media Psychology*, Volume 3, Number 1.

Hennig-Thurau, T., Gwinner, K., Walsh, G., Gremler, D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?. *Journal of Interactive Marketing. 18. 38 - 52.*

Hoßfeld, T., Schatz, R., & Egger, S. (2011). SOS: The MOS is not enough! *QoMEX 2011*, September, Belgium.

Koh, N., Hu, N., & Clemons, E. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications, 9*.

Kostakos, V. (2009). Is the crowd's wisdom biased? A quantitative analysis of three online communities. 2009 International Conference on Computational Science and Engineering, Vancouver, BC, 2009, pp. 251-255.

Liu, W. (2013). Word of mouth for movies: its dynamics and impact on box office revenue. Journal of Marketing, Vol. 70, No. 3, pp. 74-89.

Mohr, I. (2007). Buzz marketing for movies. *Business Horizons*, 50, 395-403.

Moon, S., Bergey, P., Iacobucci, D. (2010). Journal of Marketing, Vol. 74, pp. 108-121.

Oliver, M., Sargent, S., Weaver, J. (1998). The impact of sex and gender role self-perception on affective reactions to different types of film. Sex Roles, Vol. 38, No. 112.

Otterbacher, J. (2012). Gender, writing and ranking in review forums: a case study of the IMDb. Knowl Inf Syst 35, 645–664.

Sundaram, D., Mitra, K., Webster, C. (1998). Word-of-mouth communications: a motivational analysis. Advances in Consumer Research, Vol. 25, pp. 527-531.

Wanderer, J. (1970). In defense of popular taste: film ratings among professionals and lay audiences. American Journal of Sociology, Vol. 76, No. 2, pp. 262-272.

Wühr, P., Lange, B., & Schwarz, S. (2017). Tears or Fears? Comparing gender stereotypes about movie preferences to actual preferences. *Frontiers in psychology, 8, 428*.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11653

Comparing Methods to Retrieve Tweets: a Sentiment Approach

Stephan Schlosser[1], Daniele Toninelli[2], Michela Cameletti[2]
[1]Center of Methods in Social Sciences, University of Göttingen, Germany, [2]Department of Management, Economics and Quantitative Methods, University of Bergamo, Italy.

Abstract

In current times Internet and social media have become almost unavoidable tools to support research and decision making processes in various fields. Nevertheless, the collection and the use of data retrieved from these sources pose different challenges. In a previous paper we compared the efficiency of three alternative methods used to retrieve geolocated tweets over an entire country (United Kingdom). One method resulted as the best compromise in terms of both the effort needed to set it and the quantity/quality of data collected. In this work we further check, in term of content, whether the three compared methods are able to produce "similar information". In particular, we aim at checking whether there are differences in the level of sentiment estimated using tweets coming from the three methods. In doing so, we take into account both a cross-section and a longitudinal perspective. Our results confirm that our current best option does not show any significant difference in the sentiment, producing scores in between the scores obtained using the two alternative methods. Thus, such a flexible and reliable method can be implemented in the data collection of geolocated tweets in other countries and for other studies based on the sentiment analysis.

Keywords: *social media data collection methods; Twitter data; sentiment analysis; social network; geographical studies.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Our society is currently producing an enormous amount of information: just Twitter is able to generate about 500 million of tweets, daily, corresponding to 8TB of data (source: https://www.omnicoreagency.com/twitter-statistics/). These types of big data represent a great opportunity, but also pose several challenges. For example, big data produced using the Internet or social networks can be used in order to support research and decision making processes in several fields. Nevertheless, issues with these data are linked to almost any phase of their "life", starting from the collection phase (e.g. how to collect geolocalized information?) up to their use (new tools are needed to deal with such a huge amount of information), and to their analysis and interpretation (e.g. the representativeness of the covered statistical units). Most of these challenges still need to be fully explored (Goonetilleke *et al*., 2014, Alabdullah *et al.*, 2018 and Morstatter *et al*., 2013).

Our current work is focused on issues linked to the first fase: the collection of social media data. In particular, in a previous work (Schlosser et al., *forth. 2020*) we started studying three alternative methods of collection of messages sent through the Twitter social network (i.e. tweets). These methods were called M1, M2 and M3. The main advantage of all three methods is that they are all able to cover an entire geographical area, United Kingdom (UK) in our case, providing us with a set of fully geolocalized tweets. Nevertheless, the three methods have substantial differences, in terms of level of effort necessary to set them up, of spatial coverage accuracy and of the "amount" of information they are able to produce. Our preliminar study confirmed that, among the three, the best option is M2, a method that reduces the effort to be set (in comparison to M3) and the arbitrarity of decisions of the researcher and problems of overlapping between areas (in comparison to M1). Moreover, M2 produces the same quantity of information (in terms of number of tweets or gigabytes) and enhances the information quality (in terms of number of unique tweets, also reducing the processing times); M2 also leads to a more accurate coverage of the geographical sub-areas studied (UK NUTS; see: https://ec.europa.eu/eurostat/web/nuts/background).

This paper wants to further check if the different settings at the base of the three methods affect the information produced from the content point of view. For this purpose, we analyze tweets collected using all three methods by means of the sentiment analysis applying two of the most widely used lexicons, i.e. AFINN (Nielsen, 2011) and Bing (Bing, 2015). Using such scores, we compare the three methods taking into account both a cross-section (sec. 3.1 and 3.2) and a longitudinal perspective (sec. 3.3). Our expectaction is that there should not be significant differences between tweets collected using the three methods, in terms of level of sentiment (globally and at the sub-area level) and of behavior over time.

Our results confirm these expectations. This further identify M2 as the best option to retrieve geolocalized tweets on a wide geographical area. The high flexibility of such a method allows

to apply it to retrieve geolocalized tweets, setting a certain level of geographical detail and fully covering any other geographical area for any type of research purposes.

2. Literature review

Several recent research projects are based on studies applied to data coming from social networks such as Twitter. These new sources of data, together with the Internet, are also used, from a practical perspective, in order to support decision processes in several fields. In some of these cases researchers require information that is fully geolocalized: this happens, for example, monitoring socio-demographic phenomena (Jashinsky *et al.*, 2014), in disaster management (de Bruijn *et al.*, 2017) or in transportation planning studies (Paule *et al.*, 2019). In this framework, one of the biggest problem is that tweets with a geographical information are just a small fraction of the total (Middleton *et al.*, 2018). Moreover, this information is not always reliable or nicely structured (Middleton *et al.*, 2018, Zheng *et al.*, 2018), mostly because self reported by users. There are currently several methods used to overtake these limits: location extraction (Ozdikis *et al*., 2017, de Bruijn *et al.*, 2017, Zheng *et al.*, 2018) or statistical models and machine learning methods are used to assign spatial coordinates to media items basing on tweets content (Zola *et al.*, 2019, Han *et al.*, 2014). Nevertheless such methods have some limitations, because, for example, the informal and unstructured form of tweets leads to low performances of natural language processing tools (Ajao *et al*., 2015).

Our research aims to overtake such a problem, suggesting one method of tweet collection able to fully cover a geographical area (UK, in our case) and providing tweets that are fully geo-localized. We already compared in a previous work (Schlosser et al., *forth. 2020*) three alternative methods. One of them, M2, resulted the best option, as it is more efficient in comparison to the other two, taking into account the effort for its setting, the "quantity" of information produced as well as the reduction of the data cleaning times. Nevertheless, taking into account the measured sentiment, is M2 also able to perform similarly to the other methods? That is, applying two different sentiment lexicons to tweets collected using all three methods, do we obtain the same sentiment level, distribution and longitudinal evolution?

3. Data and method

In this paper we analyze all tweets collected using our three alternative methods (M1, M2, M3) in the period from January 15 to February 15, 2019. The three methods of collection and their main features are fully introduced in Schlosser *et al*. (2019) and in Schlosser *et al*. (*forthcoming*, 2020). In total, we analyzed 36,348,292 tweets for M1, 40,330,747 for M2 and 34,506,190 form M3, for a total of 111,185,229 tweets.

After first standard steps of cleaning (e.g. removing stop words and special characters and converting the text to lower case), to each tweet collected we apply both the AFINN and the Bing lexicon in order to estimate the level of sentiment. In particular the AFINN lexicon is very widely used for sentiment analysis. Its current version (AFINN-en-165) includes over 3,300 words, each of them associated to an integer score ranging from -5 (very negative) to +5 (very positive). Basing on this, we assign a score to the words of a tweet included in the lexicon and we obtain (summing up such scores) what we define as AFINN score for the considered tweet. The Bing lexicon includes 6,788 words that are classified as positive (we assign to them a value equal to +1) or negative (we assign to them a score equal to -1). The Bing score for a tweet is obtained summing up all the scores linked to the words included in it. Thus, for each tweet we analyzed we obtain two scores (an AFINN and a Bing score).

Our objective is to detect whether there are differences in the level of sentiment score detected using tweets coming from the three different methods. This is done considering three criteria, each of them applied to tweets processed by each of the two lexicons (Bing and AFINN). First, we compare the global averages and the averages by sub-areas (NUTS-1 for UK) computed on all tweets sentiment scores collected by a certain method (sec. 3.1); this is done because the sentiment is a phenomenon strongly varying, at the local level. Second, we compare the distribution of scores (sec. 3.2) to check if the method of collection affects the scores distribution. Third, we check whether our three methods perform similarly in producing sentiment estimates taking into account a longitudinal perspective, i.e. analyzing the evolution of measured sentiment by method and by day (sec. 3.3). This because in studying the sentiment, it is also (or even more) important to detect if the method of collection is able to reproduce accurately the sentiment trend and point-by-point changes over time.

4. Findings

In this section we show the main results of our analysis, according to the three criteria introduced at the end of the previous section.

4.1. Sentiment score comparison

Generally speaking, that is working on all tweets collected using the three studied methods, we did not find any statistically significant difference between the mean scores observed by method. Using the AFINN lexicon, the average score (see last row of Table 1) is equal to 0.769 for M2, to 0.758 for M1 (-1.43%) and to 0.779 for M3 (+2.77%). The average Bing score is equal to 0.247 for M2, to 0.242 for M1 (+2,14%) and to 0.251 for M3 (+3.79%). As a consequence, we confirm that M2 produces results that are intermediate in comparison to the slight underestimation of the sentiment obtained with M2 and the small overestimation obtained with M3. Nevertheless, we applied Kolmogorov-Smirnov tests to verify if there are differences in the distributions of the scores. Both at the level of individual NUTS and at the

country level no significant differences were found among the three methods. By studying the *p*-values we can conclude that the distributions obtained with the three methods are not significantly different (AFINN: M1 vs M2: $p < .001$; M1 vs M3: $p < .001$; M2 vs M3: $p < .001$. Bing: M1 vs M2: $p < .001$; M1 vs M3: $p < .001$; M2 vs M3: $p < .001$)

These very similar results can be caused by the fact that the sentiment observed on such a huge number of tweets and on such a big geographical area (UK) can be driven by a very wide range of topics of different types (pollution, economic scenery, politics, …) that define the current "mood" of Twitter user. Thus, a more reliable analysis was developed working on the Bing and AFINN scores at the level of NUTS (see the first part of Table 1).

Table 1. Average Bing and AFINN scores by NUTS and by method (M1, M2, M3).

	Bing			**AFINN**		
NUTS	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
UKC	0,255	0,246	0,276	0,807	0,771	0,864
UKD	0,226	0,227	0,241	0,708	0,711	0,755
UKE	0,250	0,263	0,244	0,791	0,829	0,777
UKF	0,228	0,226	0,220	0,700	0,702	0,682
UKG	0,251	0,253	0,247	0,764	0,764	0,751
UKH	0,291	0,281	0,289	0,886	0,856	0,885
UKI	0,201	0,185	0,188	0,649	0,598	0,608
UKJ	0,258	0,271	0,273	0,793	0,822	0,827
UKK	0,299	0,300	0,304	0,904	0,907	0,918
UKL	0,281	0,284	0,281	0,852	0,859	0,850
UKM	0,211	0,215	0,209	0,705	0,710	0,686
ALL	**0,242**	**0,247**	**0,251**	**0,758**	**0,769**	**0,779**

Table 1 shows that for both lexicons (Bing and AFINN) there are no distributional differences of the scores between the level of individual NUTS (none of the p values is above 0.01).

4.2. Distribution comparison

In Figure 1 we plot the distribution of, respectively, the average Bing and the average AFINN scores obtained analyzing all the tweets by method. Observing both figures it is easy to notice that there are no differences in the distribution by score computed with the three different methods, both in correspondence to the peaks and along the tail of the distributions.

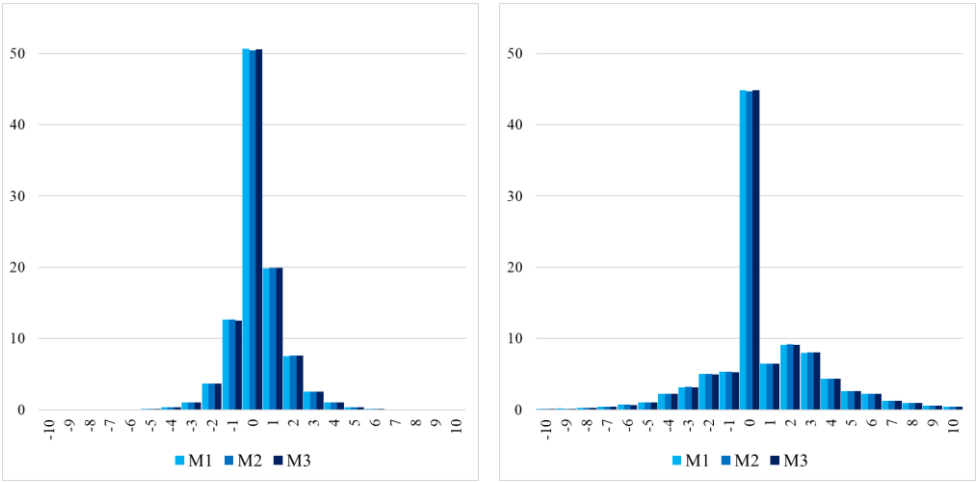


Figure 1. Distribution of the average Bing (left) and AFINN (right) scores by method (all tweets; percentage).

For both distributions we notice high central peaks corresponding to the neutral level. Moreover we notice an higher concentration around the neutral level for Bing lexicon (probably a consequence of the lower variability of the Bing scores assigned to single words) and heavier tails for the AFINN distribution.

4.3. Longitudinal analysis

In order to study potential differences between the three methods of collection in terms of content, it is also relevant to evaluate how the level of sentiment changes over time. This because research can be focused mostly on studying this feature (for a topic, regarding a theme or in a context) rather than in providing a picture referred to a specific time unit. Figure 2, referred to the AFINN lexicon, shows how the level of sentiment changes over the studied month using tweets retrieved by each of the three methods. Graphically, we observe that the three time series are very similar, showing a maximum difference of 4.7% on 2019-02-14 and a minimum difference of 0.7% on 2019-02-11, with a good overlap between the paths of broken lines representing the three methods. The results about the Bing scores are not presented here, because they confirm the findings obtained for the AFINN lexicon (shown in Figure 2). We also computed the correlation between the relative day-by-day changes among the three methods. These correlations are very high and all significantly different from zero (M1 vs M2: $r = .997$, $p < .001$; M1 vs M3: $r = .986$, $p < .001$; M2 vs M3: $r = .989$, $p < .001$).

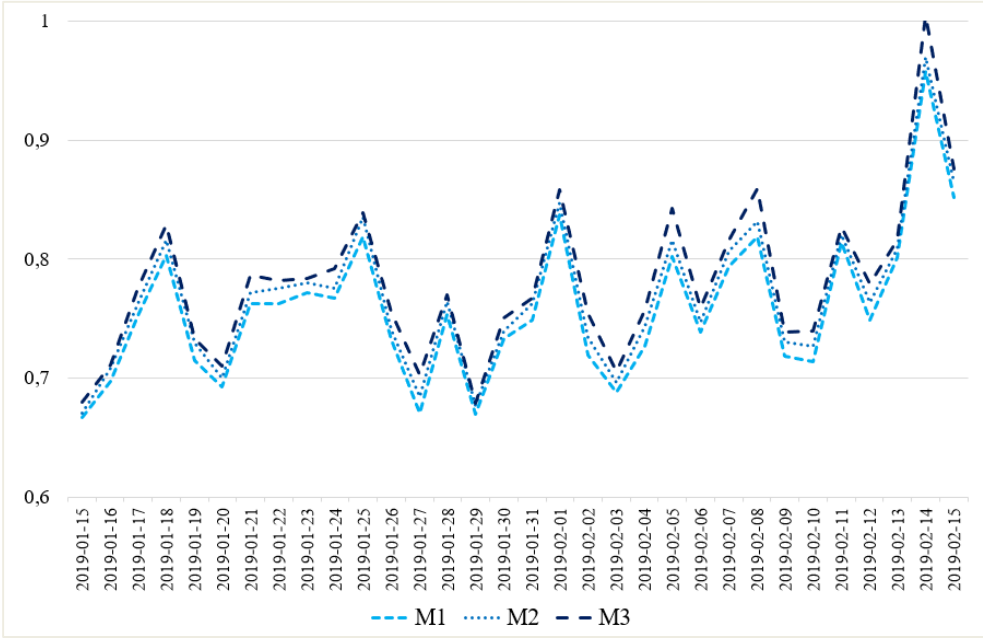


Figure 2. Daily averages of AFINN score by method (Jan. 15 to Feb. 15, 2019).

5. Discussion

In this paper we compare the performances of three methods for retrieving tweets in terms of contents, applying a sentiment analysis by using two of the most common lexicon: AFINN and Bing. Analysing the sentiment at the global level, we notice that M2 produces average scores intermediate between the corresponding scores obtaine using tweet retrieved through M1 and M3. Nevertheless, at the local (NUTS) level this does not hold for all the sub areas. However all the average sentiment scores are not significantly different; thus the three methods can be considered equivalent. If we take a look to the distribution of sentiment scores, we notice two different types of distribution for Bing and AFINN; nevertheless, there are no significant differences between the distributions obtained on tweets retrived by using the three methods. This is further confirmed by the longitudinal analysis (i.e. observing the score daily time series): the relative day by day changes show very small differences between the three methods.

As a final comment, we confirm that M2 is performing very similarly to the alternative methods (and generally produces intermediate results), considering the tweets content. Thus, we can conclude that these results further support our previous findings and that M2, also considering its flexibility, results as the best option in retrieving tweets.

References

Ajao, D., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, 41, 855–864.

Alabdullah, B., Beloff, N., & White, M. (2018). Rise of Big Data - Issues and Challenges. *Proceedings of the 21st Saudi Comput. Soc. Natl. Comput. Conf. NCC 2018*, 0–5.

Bing, L. (2015). Sentiment analysis and opinion mining. New York: Cambridge University Press.

de Bruijn, J., de Moel, H., Jongman, B., Wagemaker, J., & Aerts, J. C. J. H. (2018). TAGGS: Grouping Tweets to Improve Global Geotagging for Disaster Response. *Journal of Geovisualization and Spatial Analysis, 2, 2*.

Goonetilleke, O, Sellis, T. K., Zhang, X., & Sathe, S. (2014). Twitter analytics: a big data management perspective. *SIGKDD Explorations,* 16(1), 11-20.

Han, B., Cook, P., & Baldwin, T. (2014). Text-based twitter user geolocation prediction. *J. Artif. Intell. Res.*, 49, 451–500.

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis*, 35, 51–59.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 400-408.

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris Y. (2018). Location Extraction from Social Media. *ACM Trans. Inf. Syst.*, 36(4), article 40.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages*, Keraklion, Crete, Greece, 93-98.

Ozdikis, O., Oğuztüzün, H., & Karagoz, P. (2017) A survey on location estimation techniques for events detected in Twitter. *Knowl. Inf. Syst*., 52(2), 291–339.

Paule, J. D. G., Sun, Y., & Moshfeghi, Y. (2019). On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Inf. Process. Manag*., 56(3), 1119-1132.

Schlosser, S., Toninelli, D., & Fabris, S. (2019). Looking for Efficient Methods to Collect and Geolocalise Tweets. *Book of Short Papers SIS2019*, Milan, Italy, 1057–1062.

Schlosser, S., Toninelli, D., & Cameletti, M. (*forthcoming,* 2020). Comparing Methods to Collect and Geolocate Tweets.

Zheng, X., Han, J., & Sun, A. (2018). A Survey of Location Prediction on Twitter. *IEEE Trans. Knowl. Data Eng*., 30, 1652–1671.

Zola, P., Cortez, P., & Carpita, M. (2019). Twitter user geolocation using web country noun searches. *Decis. Support Syst.,* 120, 50–59.

Donald Trump, investor attention and financial markets

Monika Gehde-Trapp,Tapas Tanmaya Mohapatra

Department of Risk Management, University of Hohenheim, Germany.

Abstract

Information attracts attention but attention is costly. Social media has been at the forefront of information dissipation due to the sheer number of users propagating information in a fast but cheap way. We look into one specific case where Donald Trump's tweets on companies have had an effect on retail investors whose only source of information is internet. We find that retail investor attention spikes as indicated by surge in Google Search Volume Index following Donald Trump's tweets, irrespective of the tone in the tweet. We also find that Trump's tweets result in retail investors selling off stock when retail investor attention is low: retail investors sell stocks, and institutional investors buy them at later date. Finally, we analyze the daily abnormal returns of the stocks following the tweets and find that attention and tone of the tweet are opposing factors when determining abnormal returns following the tweet.

Keywords: *Investor's attention; Twitter; Retail Investors; Trading; Google SVI; Donald Trump.*

1. Introduction

When information is released, it ought to be quickly incorporated into the asset prices in efficient markets. A necessary condition for this process is attention: only when investors pay attention, newly released information can be incorporated into the asset price. However, retail investors in particular can only give attention to a few stocks in the equity universe at a given point in time since attention is costly. Kahneman (1973) was the first to raise the issue of limited attention, and Barber and Odean (2008) discuss how ignoring "right" information and paying attention to "wrong" information leads to suboptimal choices.

Investor attention can be broadly divided into retail investor attention and institutional investor attention. Institutional investors have under their arsenal a vast channel of resources to investigate stocks (e.g., Ben-Rephael et al. (2017)). In contrast, this paper focuses on retail investor attention triggered through social media, and studies its effect on equity prices. Related studies use either indirect or direct proxies of retail investor attention. Indirect proxies include absolute 1-day returns (Barber and Odean (2008)), DOW highs (Yuan (2015)) , trading volume (Gervais et al. (2001)), advertising expenses (Grullon et al. (2004)), the frequency of newspaper articles on a stock (Fang and Peress (2009)) or the appearance of a company in the New York Times (Yuan (2015)). One challenge of indirect proxies is that it is difficult to argue causally. E.g., does high trading volume cause attention, or does attention cause high trading volume? To counter this, recent studies use direct proxies for attention. These include the number of times investors login to their trading account (Sicherman et al. (2016)), the activity of investors in a brokerage account data set (Gargano and Rossi (2018)), Google search volume (SVI, Drake et al. (2012) and Vozlyublennaia (2014)), abnormal Google search volume (ASVI, Da et al. (2011)), and the Baidu index (Zhang and Wang (2015)).

So, what triggers investor attention for a particular stock? There is general consensus that media are particularly responsible for triggering investor attention. Busse and Green (2002) show that investors pay attention to morning television programs, and trade accordingly later in the day. Gurun and Butler (2012) investigate the slant of local newspapers in U.S for the local firms to satisfy the local readers. Information via social media spreads very quickly and widely, which differentiates the medium from conventional dispersal methods. Among social media platforms, Twitter is arguably among the most successful. Not only traders and important investors regularly discuss ideas and stock picks. Also, companies which are more active on Twitter have lower information asymmetry (Blankespoor et al. (2014)).

Besides this firm-initiated information, the impact of prominent figures on social media seems to be very powerful. Anger and Kittl (2011) refer to people who possess this power as "super hubs,

influencers or alpha users", representing a minority of users whose communication via Twitter reaches a widely spread and alert audiences. In this regard, the tweets of President of USA, Donald Trump plays a unique role. As of 25th February 2019, he has 58.6 million followers on Twitter. Apart from commenting on political events, President Trump also focuses on companies: Between December 2016 and January 2018, he submitted more than 50 tweets on various companies. The tone of these tweets ranges from quite harsh (Nordstrom) to extremely supportive and encouraging (Ford).

The important question then arises that how do the investors react to them? Do the tweets trigger investor attention, especially the retail investor attention? We answer these questions by using the ASVI as measure for attention. Any spike in ASVI will correspond to investor attention getting triggered. In Figure 1, we plot the average ASVI on the companies Donald Trump tweets. The x-axis denotes -15 to +15 days from the day of the tweet (0). The left panel gives the results for tweets with a negative tone, the right panel for tweets with a positive tone. We use the categories “All Categories” and “Financial Market” as our categories in Google Trend, and look for either the company name or the stock ticker. The main focus in our paper is the combination of “All Categories” and stock ticker, depicted in the blue dotted line. We find that Donald Trump’s tweets cause a significant spike in investor attention irrespective of the tone in the tweet. The attention level remains high for the day of the tweet and the following day. As expected, we find that negative tweets (left in Figure 1) create far higher attention than positive ones (right in Figure 1).

Once it is established Trump’s tweets have a strong effect on attention, we analyze subsequent trading behavior. Barber and Odean (2008) find that attention affects the retail investors more than the institutional investors. Also, it has greater impact in inducing the investors to buy rather than sell. Da et al. (2011) find that an increase in SVI leads to increased orders and trading volume by retail individual traders. In contrast, we only find an effect of negative tweets, which causes retail investors to sell off their holdings. This effect is stronger when attention prior to the tweet is already low, which is in line with Barber and Odean (2008). However, the effect we document is more long-lived and spreads out over several days. Finally, we look how the returns for the stock behave post the spike in attention and given the trading behavior of various market participants. Da et al. (2011) find that more internet search on the company lead to more upwards price pressure for the following 2 weeks.

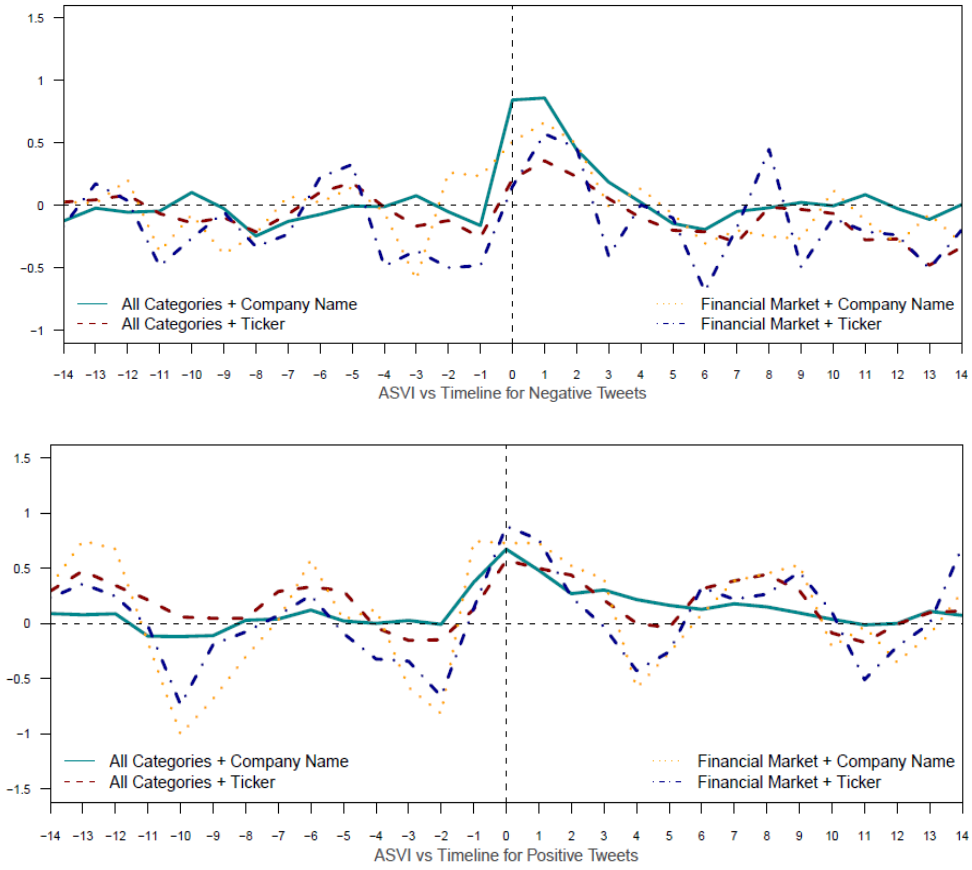


Figure 1. Average ASVI vs the days from Donald Trump's tweets.

Our study contributes to the existing literature in multiple ways. First, it studies the effect of social media on investor attention and the resulting trading behavior. Second, the paper provides evidence on the differential effect of tone on attention. Third, it provides guidance to retail investors on the detrimental wealth effects of herding due to non-fundamental information.

The rest of the paper is organized as follows. In section 2, we discuss the data sources and variables created from these sources, as well as the methodology. In section 3, we show that President Trump's tweets grab attention and lead to different trading reactions based on prior

attention. We also describe the tweets' impact on stock returns. Finally, we conclude in Chapter 4.

2. Data and Research Design

We collect all messages of President Trump from Twitter between December 2016 and January 2018. We identify company-related tweets, and assign a positive or negative tone identifier (manually). This leaves us with 45 tweets, out of which 28 have a positive and 17 have a negative tone. Next, we collect the stock tickers of the companies on which the tweet was made, and Daily Google Search Volume for these tickers from Google Trends. Tick data comes from the NYSE Trade and Quote (TAQ) database via Wharton Research Data Services (WRDS). Daily price and volume data come from Thomson Reuters Datastream.

We first test whether there is a surge in attention following Donald Trump's tweets on the companies. We derive ASVI from SVI of the stock tickers as in Da et al. (2011):

$$\begin{aligned} & ASVI_t \\ &= log(SVI_t) \\ &- log[Med(SVI_{t-1}, \dots, SVI_{t-56})]. \end{aligned} \quad (1)$$

We then test the effect of tweets on attention through the following pooled regression model:

$$\begin{aligned} & ASVI_{t,i} \\ &= \alpha + \beta * D_{t,i} + \gamma * CV_{t,i} + \varepsilon, \end{aligned} \quad (2)$$

where $ASVI_{t,i}$ is the ASVI measure for the tweet on company i on day t after the tweet. $D_{t,i}$ is a dummy variable which takes on a value of 1 on the day t for the tweet for company i, and 0 otherwise. $CV_{t,i}$ are control variables and include log of market capitalization, number of analysts followed and dollar turnover of the stock. Second, we test for the impact of the tweet on buy-sell imbalance for different investor groups (retail and institutional). Buy-sell imbalance captures the buying or selling pressure for these groups, since a negative buy-sell imbalance suggests selling pressure by a particular class of investors. To measure this effect, we create a $Order_{t,i,j}$ variable given by:

$$\begin{aligned} & Order_{t,i,j} \\ &= \frac{B_{t,i,j} - S_{t,i,j}}{B_{t,i,j} + S_{t,i,j}} \end{aligned} \quad (3)$$

$B_{t,i,j}(S_{t,i,j})$ is the buy (sell) initiated dollar volume for the company i on day t for the trader class j. We calculate the buy or sell initiated trade following the Lee and Ready (1991) algorithm. Referring to Lee and Radhakrishna (2000), we define our trader classes as small, medium, and large. We then test the effect of Trump's tweets on buy-sell imbalance via the following pooled regression model:

$$Order_{t,ij} = \alpha + \beta_1 * Order_{t-1,i,j} + \beta_2 * D_{t,i} + \varepsilon \quad (4)$$

where $Order_{t,i,j}$is the buy-sell imbalance for day t , the tweet on company i and trader class j. $D_{t,i}$ is as in equation (2). Third, we look at abnormal daily stock returns. We calculate abnormal daily return as in Zhang et al. (2016), and estimate model (5) with abnormal returns via the following pooled regression model:

$$AR_{t,i} = \alpha + \beta_1 * D_{t,i} + \varepsilon \quad (5)$$

where $AR_{t,i}$is the abnormal return for day t for the tweet on company i. We re-run the regressions separately for tweets with positive and negative tone. Also, to check the effect of attention, we additionally separate the data into top half and bottom half based on the attention the stock receives on the day of the tweet, and run separate regressions for the resulting sub-samples.

3. Results

We now look into the effect of the tweets on retail investors' attention.

3.1. Tweets and Attention

The regression analysis is done for the tweets as described in (2). As can be seen in Table 1, the coefficient for the day coefficient T0 (day of the tweet) and T1 (one day after the tweet) is positive and statistically significant in all cases. Hence, tweets increase attention for the stock. logMktCap, logTurnOver and logAnalysts are the log of market capitalization, dollar turnover and number of analysts followed for the stocks respectively.

Table 1: Dependent variable: ASVI "All categories" and stock tickers.

	All tone	Positive tone	Negative tone
T0	0.401***	0.327***	0.517***
T1	0.357***	0.339***	0.376**
T2	0.060	0.019	0.122
T3	0.034	0.032	0.016
T4	0.146	0.116	0.183
T5	0.079	0.085	0.054
logMktCap	-0.091***	-0.113***	-0.133***
logTurnover	0.058	0.013	0.169**
logAnalysts	-0.090	0.010	-0.295**
Observations	720	464	256
R2	0.069	0.065	0.129
Adjusted R2	0.057	0.047	0.097
Res. Std. Error	0.543 (df = 710)	0.539 (df = 454)	0.529 (df = 246)
F Statistic	5.859*** (df = 9; 710)	3.520*** (df = 9; 454)	4.039*** (df = 9; 246)

Note: *p<0.1; **p<0.05; ***p<0.01

This is in line with Drake et al. (2012) who find a surge in ASVI on the day of the event and post-event day. Comparing the differences between the columns, we observe that negative tweets have a 50% higher impact on attention than positive ones. Investors' attention is thus more drawn when the tone of the tweets is negative than when it is positive. The impact of the control variables is as expected: larger companies and those covered by more analysts are more transparent, resulting in overall lower search volume. High attention goes to those stocks for which investors might have to spend some effort, smaller ones followed by a small number of analysts. Turnover is positively associated with attention, but only for negative tweets.

3.2. Attention and Buy-Sell Imbalance

We now explore the impact of the tweets on buy-sell imbalance via the attention channel. We only focus on negative tweets, because the results are significant. We run the regression separately for all three trader types (small, medium, and large), and for tweets in high and low attention environments. We define a low attention environment by a below-median ASVI for the company on the day of the tweet (bottom), and a high attention environment by an above-median ASVI for the company on the day of the tweet (top). Table 2 shows the estimation results for equation (4). prevI1, prevI2 and prevI3 are the buy-sell imbalance for small, medium and large traders respectrively for 1 day before the tweet. T0 is the dummy variable which is 1 for the day of the tweet and 0 otherwise. T1 is the dummy variable which is 1 for the day of the tweet and 0 otherwise and so on.

Table 2: Dependent variable: Buy-Sell imbalance for different trader groups, following negative tweets.

	Small-Bottom	Medium-Bottom	Large-Bottom	Small-Top	Medium-Top	Large-Top
prevI1	0.575***			0.788***		
prevI2		0.361***			0.655***	
prevI3			-0.049			0.058
T0	-0.023	-0.035	0.144	-0.051	-0.066	-0.036
T1	-0.060*	-0.071	-0.070	-0.024	-0.021	-0.084
T2	-0.023	-0.019	-0.024	0.024	0.007	0.092
T3	-0.061*	-0.051	0.137	-0.024	0.001	0.071
T4	-0.016	0.046	0.113	-0.039	-0.027	-0.064
T5	0.031	-0.051	0.124	-0.030	0.010	0.136
T6	-0.091**	-0.022	0.440***	0.013	0.049	-0.032
T7	-0.071**	-0.104**	0.040	-0.031	-0.055	0.024
T8	0.025	-0.011	-0.004	0.031	0.067	-0.028
T9	-0.007	-0.012	0.555***	0.010	0.028	0.166
T10	-0.021	-0.032	0.121	-0.040	-0.014	-0.079

Note: *p<0.1; **p<0.05; ***p<0.01

For high attention stocks, buy-sell imbalance is not affected by tweets. This is interesting, since attention usually creates buying pressure. One possible explanation is the negative tone, which

may counteract the attention-based buying pressure (Tetlock (2007)). However, when repeating the analysis for positive tweets, we do not find a positive impact either. In contrast, Table 2 shows that small and medium traders increase their selling pressure following the tweets in low attention environments. At the same time, there is buying pressure from institutional investors: large traders move in to buy the stock.

Turning towards the economic interpretation of the results of Table 2, we find that small traders sell off stocks of companies following a tweet, whereas large traders buy in (for negative tweets in a low-attention environment). This result is in line with Barber and Odean (2008), who find that retail traders sell and large traders buy stocks on low attention days. We observe a staggered introduction of this pattern: Large traders strategically defer their trades. Apparently, institutional investors trade more as a reaction to retail traders' behavior.

3.3. Attention and Daily Returns

Last, we analyze abnormal returns. We run regression with daily abnormal returns as our dependent variable and the dummy day variables as the independent variable. We focus on positive tweets first, and find a positive and significant abnormal return of around 0.2% for 2, 3 and 6 days post the tweet. As the market incorporates the tweets, investors start purchasing stocks. In contrast, negative news seems to have no price impact. As in Table 2, we separate the sample into a high and low attention environment subsample based on the ASVI on the day of the tweet. The results for the tweets are in Table 3 after running equation (5). The left panel gives the results for tweets with a negative tone, the right panel for tweets with a positive tone. T0 is the dummy variable which is 1 for the day of the tweet and 0 otherwise. T1 is the dummy variable which is

Table 3: Dependent variable: Daily Returns, following positive tweets.

	Negative Tweets			Positive Tweets	
	Top ASVI Tweets	Bottom ASVI Tweets		Top ASVI Tweets	Bottom ASVI Tweets
T0	0.003	0.0004	T0	-0.002	-0.001
T1	-0.0002	0.003*	T1	0.001	0.0003
T2	0.002	0.001	T2	0.003**	0.001
T3	-0.001	0.001	T3	0.003**	0.001
T4	0.0003	0.001	T4	0.003**	-0.002*
T5	0.001	0.001	T5	0.002*	0.0003
T6	0.003	-0.001	T6	0.001	0.002**
T7	-0.002	-0.001	T7	-0.001	0.0001
T8	0.001	0.001	T8	0.001	-0.0001
T9	-0.001	-0.001	T9	0.0004	-0.001
T10	0.001	-0.0003	T10	-0.001	-0.0002
Observations	128	128	Observations	224	224
R2	0.073	0.050	R2	0.102	0.047
Adjusted R2	-0.015	-0.040	Adjusted R2	0.055	-0.002
Res Std. Err. (df = 116)	0.005	0.004	Res. Std. Err (df = 212)	0.004	0.004
F Stat (df = 11; 116)	0.825	0.555	F Stat(df = 11; 212)	2.181**	0.960

Note: *p<0.1; **p<0.05; ***p<0.01

For the positive tweets, from 2 to 5 days after the tweet, stocks show an abnormal return of 0.2% to 0.3% in the high attention environment. For the low attention environment, we obtain a significant return of around -0.2% 4 days after the tweet, which reverses on day 6 post the tweet. Next we repeat the exercise for negative tweets. We only get a +0.3% abnormal return for the first day after the tweet in the low attention environment. There is no significant abnormal return

for the high attention environment. This may be due to the offsetting effects of attention and tone: negative tone should result in a negative return (Tetlock (2007)), but high attention should result in positive returns (Barber and Odean (2008)). Both the phenomena seem to act in opposite to each other resulting in almost no change in the abnormal returns for the stocks.

4. Conclusion

Attention is costly and retail investors react differently once their attention is grabbed. In recent years, social media has played an important role in grabbing investor attention for stocks. News about a stock dissipate fast and cheap thus reaching a wide audience. In this regard, we check for a particular activity which is gathering investor attention through social media: Company-related tweets by President Trump. We find that his tweets cause a significant spike in attention on and directly after the day of the tweet. Tweets with a negative tone create 50% more attention than tweets with a positive tone.

We then analyze how different trader groups react to the trades. In line with the lower attention, positive tweets do not affect buy-sell imbalance. Negative tweets, however, result in retail investors selling off, and institutional investors buying in later days. The effect is stronger in low attention environments. Our study is the first to document this effect: the selling pressure created through the negative tone dominates the (hypothesized) buying pressure through increased attention. Institutional investors take advantage of this behavior by retail investors, and buy up the stocks that retail investors sell. As a result, retail investors lose out to institutional investors. It can be seen that while retail investors are selling during negative tweets, there is slight price rise on 1 day after the tweet and no fall after that. It is important therefore that retail investors maintain caution when President Trump tweets about a given company.

Similarly, for positive tweets, when checked through the window of attention for days following the tweet, high ASVI stocks show good abnormal returns whereas low ASVI positive tweet stocks show no significant abnormal returns. This is in line with existing literature that high attention results in abnormal returns. Negative tweets on a whole do not show any significant returns post the tweet. Thus it is important that the retail investors do not sell off their stocks post the tweet if there is negative news about a company because there is no significant fall in stock prices which happen after the tweet. It is also important to analyze as to why some stocks get higher attention than the others to make more incisive analysis.

Acknowledgments

We thank the participants of the 5th Winter Finance Workshop of the University of Hohenheim 2018 for helpful comments. We gratefully acknowledge access to the Center for Research in Security Prices (CRSP) and DataStream provided by DALAHO, University of Hohenheim.

References

Anger, I.; Kittl, C. (2011): Measuring Influence on Twitter. In S.N. Lindstaedt, M. Granitzer (Eds.), *I-KNOW 2011,11th international conference on knowledge management and knowledge technologies, graz, austria, september 7-9,2011.*

Aouadi, A.; Arouri, M.; Teulon, F. (2013): Investor attention and stock market activity. Evidence from France. In *Economic Modelling*, 35, 674–681.

Barber, B. M.; Odean, T. (2008): All That Glitters. The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. In *The Review of Financial Studies*, 21(2), 785–818.

Ben-Rephael, A.; Da, Z.; Israelsen, R. D. (2017): It Depends on Where You Search. Institutional Investor Attention and Underreaction to News. In *The Review of Financial Studies*, 30(9), 3009–3047.

Blankespoor, E.; Miller, G. S.; White, H. D. (2014): The Role of Dissemination in Market Liquidity: Evidence from Firms' Use of Twitter. In *The Accounting Review*, 89(1), 79–112.

Busse, J. A.; Green, T. C. (2002): Market efficiency in real time. In *Journal of Financial Economics*, 65(3), 415–437.

Da, Z.; Engelberg, J.; Gao, P. (2011): In Search of Attention. In *The Journal of Finance*, 66(5), 1461–1499.

Drake, M. S.; Roulstone, D. T.; Thornock, J. R. (2012): Investor Information Demand. Evidence from Google Searches Around Earnings Announcements. In *Journal of Accounting Research*, 50(4), 1001–1040.

Fang, L.; Peress, J. (2009): Media Coverage and the Cross-section of Stock Returns. In *The Journal of Finance*, 64 (5), 2023–2052.

Gervais, S.; Kaniel, R.; Mingelgrin, D. H. (2001): The High-Volume Return Premium. In *The Journal of Finance*, 56(3), 877–919.

Grullon, G.; Kanatas, G.; Weston, J. P. (2004): Advertising, Breadth of Ownership, and Liquidity. In *The Review of Financial Studies*,17(2), 439–461.

Gurun, U. G.; Butler, A. W. (2012): Don't Believe the Hype: Local Media Slant, Local Advertising, and Firm Value. In *The Journal of Finance*, 67(2), 561–598.

Kahneman, D. (1973): Attention and Effort: Englewood Cliffs,NJ,Prentice-Hall.

Lee, C.M.C.; Radhakrishna, B. (2000): Inferring investor behavior: Evidence from TORQ data. In *Journal of Financial Markets*, 3(2), 83–111.

Lee, C.M.C.; Ready, M. J. (1991): Inferring Trade Direction from Intraday Data. In *The Journal of Finance*, 46(2), 733–746.

Sicherman, N.; Loewenstein, G.; Seppi, D. J.; Utkus, S. P. (2016): Financial Attention. In *The Review of Financial Studies*, 29(4), 863–897.

Tetlock, P. C. (2007): Giving Content to Investor Sentiment. The Role of Media in the Stock Market. In *The Journal of Finance*, 62(3), 1139–1168.

Vozlyublennaia, N. (2014): Investor attention, index performance, and return predictability. In *Journal of Banking & Finance*, 41,17–35.

Yuan, Y. (2015): Market-wide attention, trading, and stock returns. In *Journal of Financial Economics*, 116(3), 548–564.

Zhang, B.; Wang, Y. (2015): Limited attention of individual investors and stock performance. Evidence from the ChiNext market. In *Economic Modelling*, 50, 94–104.

Zhang, Y.; Song, W.; Shen, D.; Zhang, W. (2016): Market reaction to internet news. Information diffusion and price pressure. In *Economic Modelling*, 56, 43–49.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11655

#immigrants project: the on-line perception of integration

Rosario D'Agata, Simona Gozzo
Departement of Political and Social Sciences, University of Catania, Italy.

Abstract

This paper analyses the content of Twitter's comments during the period covering the last European elections. "#immigrants" is the extraction's keyword in different national languages. With the exception of English and French, whose extraction would be misleading, all of the other languages have been chosen to catch the geographical area of reference. We made sure to extract at least two sentences for each Welfare area. Once the data have been extracted, three different strategies have been used. The first one, dealing with both a qualitative and a quantitative assessment; the second one, analysing automatically the content of the top 10 extracted tweets during the reference period and the third one based on network analysis. Through a deep analysis of the content, three clusters have been identified: the first one dealing with the cultural risks of multiculturalism; the second one (social risks) dealing with the fear of migrants stealing job vacancies and the third one dealing with economic risks. A deep network analysis of Italian and Spanish contexts follows. What emerges is that: communication is extremely heterogeneous; in Italy there unique and duplicated edges prevails; in Spain there are more groups than in Italy, more themes covered and different kind of users and nets.

Keywords: *Big data; immigration; Network analysis; twitter.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

This work concerns a study carried out by analysing the content of Twitter's comments during the two months close to the last European elections. The extracted tweets contain the headword "#immigrants" in different national languages. The selection of this specific social network platform derives from the decision to identify an instrument oriented to specific forms of communication: that carried by high interest in politics and/or involvement (Laifeld 2018). The detected discussions are conveyed not only to express an opinion, but also to influence public opinion with regard to the issue of integration.

A constraint of the approach implies the exclusion of comments in French and English. This depends on the need to attribute - for the interpretation of the comments - the sentences in a language to a reference area.

The analysis following the extraction of data follows three different comparing strategies allowing to detect complementary information. At the same time, the goal of this study is to assess when the three approaches capture redundant information and when, on the other hand, it is useful to integrate them. In particular, the applied strategies are:

1) A study based on an in-depth qualitative analysis (content) and quantitative (number of underlying links and ego-network structure) assessment. This analysis is centred on the top 10 extracted tweets in the selected language unit and for each week. The number of tweets in Polish and Slovenian languages is very limited. Probably this is due to the preference of other social networks (Surowiec and Štětka 2018).

2) Automatic analysis of the content of top 10 tweets extracted during the reference period

3) A study based on network analysis tools. This choice allows to select – for each week – the whole structure of relational net and the main groups as sub-networks obtained by extracting clusters mutually connected, with higher internal homogeneity and external heterogeneity in terms of links. This stage is referred specifically to Italian and Spanish comments.

2. The on-line comments

The period during which online communication was monitored (weekly) is close to the 2019 European consultations: from April 15 until early June 2019. This extraction is made in order to include mobilization effects on specific elements and / or priorities in communications during the last part of May. The increase in communication in this period could represent the political choice to convey the electoral campaign on issues aimed at either promoting integration or exclusion. This emerges as a strong trait for two Scandinavian areas (Sweden and Denmark), while an increase in communication - although not particularly high

compared to the dynamics that emerged - is recorded in the Mediterranean areas. No effect arises Eastern Europe areas.

What emerges immediately, considering the total number of users for each week, is that the Eastern Europe areas have no bias for communicating via Twitter. Communication in Polish and Slovenian, in fact, involved a few dozen contacts for week.

Table. 1. Number of users within the nets and for the top 10 tweets selected (page-rank) for each week.

Data	DK	FI	DE	IT	NO	NL	PL	PT	SI	ES	SE
15-22 april	528	525	985	9636	1472	1472	15	np	24	17023	2583
23-29 april	492	210	755	11736	1397	1397	26	np	21	18457	2157
30 april - 6 may	512	302	860	9139	1832	1832	42	4176	38	14392	2552
7-13 may	422	279	1129	10348	3093	3093	81	3606	np	17788	2347
14-19 may	371	239	845	11009	2732	2732	61	3189	42	np	2719
20-26 may	739	323	1172	10330	1744	1744	57	4724	13	18189	3108
27 may - 3 june	442	204	879	10792	2105	2105	np	3925	29	17628	3322
Node for top 10 tweets (%)											
15-22 april	33.71	40.38	42.34	34.63	52.51	31.39	73.33	np	58.43	58.43	52.42
23-29 april	36.18	28.10	11.13	41.14	22.96	23.19	65.38	np	30.96	30.96	41.68
30 april - 6 may	29.10	25.83	18.95	30.23	23.28	36.3	52.38	np	47.08	47.08	32.41
7-13 may	19.43	32.97	39.50	33.00	27.76	4.95	2.47	37.02	46.67	46.67	42.05
14-19 may	22.37	28.87	31.12	25.33	22.76	47.25	24.59	21.16	np	np	37073
20-26 may	49.12	45.51	42.41	2.91	20	29.87	29.82	14.39	23.02	23.02	37.19
27 may - 3 june	28.96	40.20	13.31	3.79	19.4	41	np	46.42	37.25	37.25	31.37
Average	31.27	34.55	28.39	24.43	26.95	30.56	41.33	29.75	40.57	40.57	39.26

The opposite trend, namely that of a particularly widespread and pervasive communication, is mainly registered in the Mediterranean Europe (Portugal, Italy and, above all, Spain) and in Sweden and Netherlands. The most relevant network comments cover only part of the entire communication flow. However, this selection permits an in-depth qualitative analysis on every single twit for each language, i.e. more specific information about the sources of information – not easy to catch (Tab. 2), the meaning of the tweet and its purpose. What has

been detected, when possible, is that news published either in online newspapers or private comments prevail, especially in Sweden, Spain and Finland, while Italy is characterized by an unusually high proportion of institutional comments by both parties and politicians (the social and web campaign conducted by the main governing parties is widespread).

Further information refers to the content of the tweets (Tab 3). All comments and posts refer to prejudices of some kind against immigrants and yet some specificities emerge quite clearly. Scandinavian areas, for example, are particularly sensitive to the issue of subsidies to immigrants, which are specifically considered a major problem in Denmark.

Table 2 - Source of information on twitter.

Country	Parties or politics	Trade Unions	Newspapers	Private users	Total
Denmark	7	0	10	5	22
Finland	3	5	7	12	27
Germany	5	2	10	2	19
Italy	21	2	1	11	35
Norway	3	0	8	3	14
Netherlands	3	0	8	11	22
Poland	0	0	1	1	2
Portugal	3	0	5	8	16
slovenia	0	0	4	2	6
Spain	6	0	2	13	21
Sweden	*3*	0	11	46	60
Tot	***54***	***9***	***67***	***114***	***244***

Tab. 3. Themes in tweets (%).

Country	Against prejjudices	Against immigrant subsidies	Prejudices	Analysis of prejudices and immigration argument	Against parallel comminities that do not communicate	Against institutions or politicians	Against journalist	Integration and work
DK	11,9	25,4	31,3	23,9	3,0	1,5	3,0	0,0
FI	11,6	2,9	24,6	47,8	2,9	5,8	0,0	4,3
DE	6,0	9,0	26,9	44,8	0,0	4,5	0,0	9,0
IT	11,8	1,5	33,8	25,0	1,5	26,5	0,0	0,0
NO	17,9	6,0	17,9	38,8	0,0	16,4	0,0	3,0
NL	1,6	8,1	32,3	32,3	0,0	24,2	0,0	1,6
PL	10,0	0,0	20,0	62,5	0,0	5,0	0,0	2,5
PT	16,0	2,0	8,0	46,0	0,0	12,0	0,0	16,0
SI	11,4	2,3	25,0	31,8	0,0	11,4	0,0	18,2
ES	13,6	4,5	31,8	18,2	2,3	25,0	4,5	0,0
SE	8,3	8,3	36,7	8,3	0,0	30,0	5,0	3,3

Moreover, unlike what is often believed, Scandinavian areas are among the main ones to express prejudices and fears, despite many comments aimed at "understanding" and solving problems related to the migration phenomenon. The most tolerant of the Scandinavian areas seems, in this sense, to be Finland, while the one less inclined to welcome and understand is Denmark. On the other hand, the Mediterranean areas are mainly characterized by institutional criticism (political, social, cultural). Concerning the subjects of the tweets (Tab. 4) we notice that *Political institutions* are present in 45.5% of Dutch tweets and *Parties* in 31.8% of Danish ones. In Spain, most of the tweets refer to *National Community.* In Germany and Portugal, almost one out of two tweets, concerns *immigrants* in general and in Sweden 1 out 5 *urban violence* related with immigrants.

At a later stage, a qualitative analysis has been carries out, analysing the 'sound' of the tweets (Tab. 5). The most positive evaluations seem to be twitted in Finland (14.5%) and Portugal (14.0%). On the contrary, the most negative ones appear in Netherlands (58.1%) and in Spain (57.8%).

Tab. 4 – Subjects of tweets (%)

	Political institutions	*Parties*	*Associations*	*National community*	*Immigrants*	*Extremists (violence)*	*Poverty -ghetto*	*Immigrants (urban violence)*
DK	13,64	31,82	0,00	9,09	28,79	6,06	3,03	7,58
FI	26,09	4,35	5,80	15,94	34,78	0,00	1,45	11,59
DE	14,93	8,96	7,46	8,96	49,25	0,00	0,00	10,45
IT	27,94	22,06	10,29	17,65	13,24	0,00	0,00	8,82
NO	11,94	22,39	5,97	22,39	20,90	1,49	1,49	13,43
NL	43,55	17,74	1,61	9,68	20,97	1,61	0,00	4,84
PL	5,00	20,00	5,00	17,50	35,00	0,00	0,00	17,50
PT	17,65	7,84	5,88	13,73	50,98	3,92	0,00	0,00
SI	28,89	11,11	0,00	22,22	33,33	0,00	0,00	4,44
ES	26,67	15,56	0,00	37,78	4,44	4,44	0,00	11,11
SE	28,33	18,33	5,00	16,67	10,00	1,67	0,00	20,00

Finally, we focused on the content of tweets related with internal of external policy. It is interesting to notice that while in Scandinavian area the content focuses on *Internal Policy* (94.5% average percentage), in Portugal 2 out 3 of tweets speak about foreign policy as well as 3 out 10 tweets relived in Germany.

Tab. 5 – Evaluations of immigrants and Content of tweets.

Country	Positive	Negative	Neutral	Internal policy	foreign policy
Denmark	1,49	41,79	56,72	96,97	3,03
Finland	14,49	24,64	60,87	98,55	1,45
Germany	4,41	23,53	72,06	70,15	29,85
Italy	7,35	41,18	51,47	85,29	14,71
Norway	5,97	29,85	64,18	97,01	2,99
Netherlands	0,00	58,06	41,94	80,65	19,35
Poland	10,26	28,21	61,54	89,74	10,26
Portugal	14,00	16,00	70,00	33,33	66,67
slovenia	0,00	37,78	62,22	82,22	17,78
Spain	11,11	57,78	31,11	86,67	13,33
Sweden	3,33	53,33	43,33	93,33	6,67

Main tweets' comments

The Clustering procedure carried out on the content of top-10 tweets selected for each week allowed identifying the main thematic cores. From that, three different clusters emerged. The results are analysed looking at the chi-square value and the frequency of each lemma within each cluster and within the tweet (defined as "elementary context"). These analyses allow to name and define the different identified clusters.

The first cluster (*cultural risks*) includes tweets aimed at questioning the multi-cultural approach. Comments from Finland and Sweden are mainly associated to this thematic group. One of the main issues is the of inability to integrate, in particular with regard to the differences in language, habits and culture. However, anti-racist comments in German and Portuguese are also related to this cluster.

The second cluster (*social risks*) is also the one in which most part of comments converges and includes many phrases, especially in Danish (but also in Polish), that use the terms as "work" and/or "people".

The third cluster (*economic risks*) is the one that is numerically less consistent and it refers to those comments that include terms such as Euro, Europe, immigrants, illegal immigrants, foreigners. This cluster is the least specific because it is associated with comments from all countries (although the association is stronger with Eastern Europe).

Tab. 7. Weight of clusters identified by elementary contexts analysis.

CLUSTER 1	341	31.31%
CLUSTER 2	455	41.78%
CLUSTER 3	293	26.91%

The clustering effects

In this step relational structures are analysed through network analysis tools (Borgatti *et al.*, 2013), reconstructing the graphs for each week, then extracting those groups with greater internal homogeneity and external heterogeneity (looking at both unique and reciprocal links). On the other hand, the frequent display of same tweets, domains and information probably implies a conformity of perspective among users. We can verify this hypothesis looking at the main content of the tweets for each sub-net. The analysis, carried out for each week, aims to assess precisely the presence of *communities* (although weak and iridescent) that share perspectives and materials. Furthermore, the analysis shows a very high heterogeneity which is reflected in the detection of a very high number of groups per week (extracted by applying the algorithm Clausette-Newman-Moore). As a discriminating criterion we agreed to take the presence of at least 1000 nodes in the network. In such a way, the reference groups get considerably reduced (Tab. 8). This choice does not produce a lack in information but avoids redundancies and "noise" due to self-referential or isolated nets. Besides, some differences in either area, contents and weeks emerge.

The above descripted procedure was mainly applied to the Italian and Spanish contexts, which demonstrated to be similar under many points of view: belonging to the Mediterranean context, lack of integration, transitional measures mainly oriented towards temporary reception of immigrants, with actions often defined as "emergency measures", mirroring in both areas a high level of prejudice towards immigrants.

Tab. 8 - Vertices for weeks in Italy and Spain (averages).

	Group	Vertices	Unique Edges	Edges With Duplicates	Total Edges	Self-loops
weeks			**SPAIN**			
First	4	1870	2178	136	2314	30
Second	3	1665	1792	56	1848	54
Third	4	1656	2336	191	2528	65
Forth	2	3346	4070	81	4151	28
Fifth	3	1443	2700	503	3203	126
Sixh	3	2096	3169	456	3625	106
Seventh	5	1587	2229	279	2508	36
Weeks			**ITALY**			
First	2	2064	3654	907	4561	295
Second	3	2298	3836	778	4614	235
Third	4	1203	1834	480	2315	143
Forth	3	1728	2883	1046	3929	257
Fifth	1	1736	3251	974	4225	679
Sixth	2	2236	4261	1162	5423	410
Seventh	3	1532	2403	845	3248	224

This analysis shows a greater network presence of Spanish users, although communication in Italian is pervasive. In addition, the peak of connected subjects who interact in Spanish is recorded during the fourth week while Italian users prevail in the second and sixth weeks. The number of subjects in the networks is higher in Spain while the number of links (especially in the duplicate proportion) is higher in Italian, as well as self-loops. The Italian nets appears, therefore, more redundant than the Spanish ones, which are more heterogeneous (more homogeneous groups of larger entities). Further information can be referred to the content of tweets. Indeed, these structural differences can probably be better understood by looking at the type of communication underlying them and, in particular, by analysing the sources (institutional, private, academic, political, etc.) and the users' goal (Yang et alii 2018). A similar operation was carried out through the qualitative analysis of the top 10

tweets. In this case, however, the selection becomes fully automated on the basis of Urls, Hashtag, words, etc. in each identified group.

Unlike the previous one, which selected the top 10 tweets of the week, the analysis proposed in this section refers to groups that emerged from the entire extracted network. The selection is made subsequently by importance of the single tweets with respect to the number of times it is selected or displayed and analysing the entity of the overall relational flows. The network structure identifies homogeneous sub-networks with respect to dynamics that are reconstructed, specifically, taking into account the structure of edges (Pfeffer, 2018). This procedure allows to analyse a very high number of nodes and links by identifying homogeneous networks.

Moreover, each group is homogeneous with respect to the content of the information conveyed within it because the links are given by the same information flows. Therefore, if a sub-network is more homogeneous, this depends on the fact that within it the nodes contact and send each other the same contents (or comments on them). In this way, it is therefore possible to establish a relationship between links, network structure, communication content and heterogeneity. What emerges considering the structures of networks is that in Italy and Spain:

- communication is extremely heterogeneous (this leads to a loss of information that is as relevant as you choose to select only the main networks, eliminating the noise from micro-communications or self-loops);
- in Italy there are more main unique edges and edges with duplicates (so there is more propensity to communication but even more redundancy in communication)
- in Spain there are more groups than in Italy, that is there are more themes, probably different kind of users and nets.

3. Conclusions

The information on the structure of networks has a relative value if it is not related to the content of the main tweets. The presence and the combination of different methods, with both analysis qualitative and quantitative, is the main novelty of the work. The qualitative step refers to messages that have a higher page-rank index in the network. The analysis of the content of the single tweets, furthermore, allowed to identify the type of senders. Finally, the use of the NA permitted to identify the structure of the links that, together with the content of the messages, allows further considerations in a more accurate comparative perspective.

References

Borgatti S., Martin E. &Johnson J. (2003), *Analysing Social Networks*, Los Angeles, Thousand Oaks, CA, London, Sage Publications.

Laifeld P. (2018), *Discourse Network Analysis. Political Debates as Dynamic Networks*, Victor J.N. & Montgomery A.H. (Eds) the Oxford Handbook of Political Networks, Oxford Un Press, 301-325.

Pfeffer J. (2018), *Visualization of Political Networks,* Victor J.N. & Montgomery A.H. (Eds) the Oxford Handbook of Political Networks, Oxford Un Press, 277-299.

Surowiec P. & Štětka V. (2018), Social Media and Politics in Central and Eastern Europe, Routledge

Yang S. & Gonzàlez-Bailòn S. (2018), *Semantic Networks and Applications in Public Opinion Research*, Victor J.N. & Montgomery A.H. (Eds) the Oxford Handbook of Political Networks, Oxford Un Press, 326-353.

Digital footprint for tourism research

Eduardo Cebrián, Josep Domenech
Department of Economics and Social Sciences, Universitat Politècnica de València, Spain.

Abstract

Tourists leave some digital footprints spread across a wide variety of repositories that can be studied to observe and analyze their behavior. This paper exhaustively analyzes the use of Big Data sources for understanding and predicting the main variables affecting tourist behavior. Analyzed sources include those derived from the tourist activity in the Internet and also some other digital footprint data not related to the Internet activity. The classification of sources is grounded in a model of purchase consumption system applied to leisure travel behavior. This model defines potential predictors on travelers' choices and classifies them in three stages: pre-trip, during-trip and post-trip. Our work classifies the digital footprints left by tourists according to the stage in the model and the variable they help predict or understand. As a result, a complete map of Big Data sources for tourism research is presented. This map evidences not only complementarities among sources, but also potential applications of digital footprint analysis that have not been studied yet.

Keywords: *Big Data; PCS; Tourism; Data Sources; Classification.*

This work has been partially supported by the Spanish National Research Agency (AEI) with the grants PID2019-107765RB-I00 and PEJ2018-003267-A-AR.

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

Editorial Universitat Politècnica de València

Predicting SME's default: some old facts and a new idea

Lisa Crosato[1], Josep Domenech [2], Caterina Liberati[3]
[1]Department of Economics, Ca' Foscari University of Venice, Italy, [2]Department of Economics and Social Sciences, Universitat Politecnica de Valencia, Spain, [3]Department of Economics Management and Statistics, University of Milano-Bicocca, Italy.

Abstract

The Small Business Act of the European Commission in 2008 acknowledges the key role of Small and Medium Enterprises (SMEs) in the EU economy. This is particularly relevant for Italy, which has the largest share of SMEs in Europe, as well as for other countries such as Portugal, Spain and Greece. On the other hand, SMEs experience more difficulties in their early stages mainly due to high market competition and credit constraints, as highlighted by Fritsch and Weyh (2006). For these reasons, the study of SMEs default risk is always relevant. There are several papers studying firm default factors in a single country (see Ciampi, 2015, Fantazzini and Figini, 2009, Flix and dos Santos, 2018). The literature concentrates mainly on financial indicators built on businesses' balance sheets, which are available about two years late with respect to their reference period. This diminishes the significance of the results, both for credit risk and policy aims, and particularly in a forecasting perspective. The purpose of this paper is to provide a preliminary study on a sample of spanish firms selected from the SABI, Sistema de Análisis de Balances Ibéricos, which is listed among Bureau van Dijk databases. The analysis will be carried out according to both parametric and non-parametric discrimination techniques, with the standard construction of a training set on which to build a model and a validation set to test the validity and robustness of the results, and, in the end, the reliability of the model in predicting default. Finally we present a new proposal: a scheme to understand to what extent firms' default can be predicted by substituting the traditional data sources (offline information) with data collected from their corporate websites (online information) in order to exploit more up-to-date information.

Keywords: *Default risk; SMEs; Web scraping; Corporate websites.*

This work has been partially supported by the Spanish National Research Agency (AEI) with the grant PID2019-107765RB-I00.

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

Journalists as end-users: quality management principles applied to the design process of news automation

Laurence Dierickx
ReSIC, Université Libre de Bruxelles, Belgium.

Abstract

ISO 9000 refers to a family of standards related to quality management. It defines the concept of quality as the features and characteristics of a product, a process, or a service that bears on its ability to satisfy needs explicitly or implicitly expressed. Standards provide guidance and tools to ensure that products or services will meet users 'requirements. It means that quality must be consistently improved and that risks must be evaluated to be anticipated. The seven principles of the ISO 9000 are here examined through the lenses of a case study conducted within a Belgian newsroom, where a news automation system was developed to support the daily routines of financial journalists. As end-users, they have been involved in the design process. Journalists provided the text templates to automate, based on financial data bought to a German company, while the development of the writing engine was taken in charge by a French start-up. The empirical material was collected through participant observation, including access to all of the working documents. Although it does not guarantee the quality of the content or the end-uses, the standards allowed us to frame a social process where all of the stakeholders were taken into account.

Keywords: *news automation; professional practices; quality management; standards; ISO 9000.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

Identification of online reviews helpfulness using Neural Networks

María Olmedilla[1], Rocío Martínez Torres[2], Sergio Toral[3]
[1]SKEME Business School, France, [2]Department of Business and Marketing, Universidad de Sevilla, Spain, [3]Department of Electronic Engineering, Universidad de Sevilla, Spain.

Abstract
During the last decade, research has shown that identifying helpful reviews from a big amount of user-generated review data has been a trending topic. This study proposes a classification system using an adaptive implementation of 1D Convolutional Neural Networks (CNNs) that can early identify whether an online review is helpful, fair or not helpful with 80% of accuracy. After using the neuronal encoding, a cluster analysis of the helpful and not helpful was made. The results reveal that the most significant words and documents for helpful reviews clusters describe cars and their characteristics. Whereas not helpful reviews clusters express details on car-related shops/companies in general.

Keywords: *helpfulness; online reviews; Convolutional Neural Networks; prediction; classification.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

User-defined Machine Learning Functions

Markus Herrmann, Marc Fiedler
Global Data Science, GfK, Germany.

Abstract

In Data Science practices it is commonly assumed and accepted to abstract and slice big data architectures into functional layers, in particular a triad of governance-, data analysis- and persistence layer. However, moving input data to analysis, which is required when abstracting a data persistence layer from a data analysis layer, needs to be considered as highly expensive at large scale. Especially in Machine Learning (ML), the data analytics layer module requires intense data movements during preprocessing, data integration, preparation and analytics steps.

Therefore, we propose to consider an application of User-defined functions (UDFs) with ML capabilities directly at the data persistence layer, i.e. at the database. We observed that it might be overall most efficient in traditional on-premise (i.e. non-cloud) RDBMS environments to apply ML UDFs if only singular and self-contained ML tasks should be integrated.

Whereas the availability of ML functions in databases was predominantly owned by proprietary solutions in the past, there are now entirely new opportunities to integrate Python ML libraries with open source RDBMS. Whilst considering Python as one dominant language for ML applications in Data Science, the now achieved facilitation of Python ML UDFs consequently opens a broad range of opportunities to add Python ML capabilities to already existing persistence layers - without having to build an additional data analysis layer and related pipeline.

With this presentation we deliver preliminary results of our industry research about database centric ML applications, and we open source code for the application of (un)supervised learning models.

Keywords: *Machine Learning Engineering; RDBMS; UDF; MLUDF.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

Internet searches as a leading indicator of house purchases in a subnational framework: the case of Spain

Concha Artola[1], Jorge Herrera de la Cruz[2]
[1]International Economics and Euro Area Department University, Bank of Spain, Spain, [2]Department of Economic Analysis and Quantitative Economis, Universidad Complutense de Madrid, Spain.

Abstract

Most people use web search tools to collect information on goods or services they intend to buy. Given the prominence of Google among the search engines and the availability of Google trends (GT) as a tool packaging some characteristics of those searches (geography, topic, categories, among others) it is only natural to use this instrument in order assess trends in the market.

In this paper we build indicators reflecting the real estate market stance. To do so we rely GT's TOPIC´s option that approximates the concept (housing, purchase, sale ...) instead of the exact wordings used by searchers. This approach is particularly useful in a country with several official languages and an important foreign market.

The baseline quarterly model describes house sales (measured by its year-on-year growth rate) as an autoregressive AR (1/4) model and unemployment rate as a covariate. The alternative augments the baseline with contemporary a Google indicator. The models are estimated for 2004Q1-2014Q4 and recursive one period ahead forecasts are made for 2015Q1-2018Q4. The inclusion of Google indicator reduces the EAM of prediction errors (outside the sample) from 0.077 to 0.034. The forecasts also have greater accuracy and lower bias. The same procedure has been replicated for regions with very similar results for the main regional markets (Madrid and Catalonia) and more unequal in other regions.

Keywords: *Housing markets; Spain; Google trends.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

Causal discovery with Point of Sales data

Peter Gmeiner
Global Data Science, GfK SE, Germany.

Abstract

GfK owns the world's largest retail panel within the tech and durable good industries. The panel consists of weekly Point of Sales (PoS) data, such as price and sales units data at store level. From PoS data and other data, GfK derives insights and indicators to generate recommendations with regards to e.g. pricing, distribution or assortment optimization of tech and durable good products. By combining PoS data and business domain knowledge, we show how causal discovery can be done by applying the method of invariant causal prediction (ICP).

Causal discovery, in essence, means to learn the actual cause and effect relations between the involved variables from data. After finding such a causal structure, one can try to further specify the function classes between those identified cause-effect pairs. Such a model could then be used to predict under intervention (predict when the underlying data generating mechanism changes) and to optimize and calculate counterfactual effects, given current and past data. In our development, we combine recent achievements in causal discovery research with PoS data structure and business domain knowledge (in the form of business rules).

The key delivery of this presentation is to show fundamental differences between a causal model and a machine learning model. We further explain the advantages of combining a causal model with a machine learning model and why causal information is key to provide explainable prescriptive analytics. Furthermore, we demonstrate how to apply ICP (for sequential data) to context-specific PoS data to achieve improved models for sales unit predictions. As a result, we obtain a model for sales units that is on the one hand derived from observed data and on the other hand driven by business knowledge. Such a refined prediction model could then be used to stabilize and support other machine learning models that can be used for generating prescriptive analytics.

Keywords: *Causal Model; Causal Discovery; Machine Learning; PoS.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

Interpretable Machine Learning - An Application Study Using the Munich Rent Index

Julia Brosig
Data Scientist at Sqooba Deutschland GmbH, Germany.

Abstract

Interpretable machine learning (IML) helps to understand decisions of black box models and thus improves confidence in machine learning models. To use interpretable machine learning methods, a black box model is fitted first, and on top of this model-agnostic interpretable machine learning methods are applied.

This paper analyses model-agnostic tools with regard to their global and local explainability. The methods are validated using a practical example of the estimation of the Munich rent index 2017.

In order to explain global decisions of the machine learning model, the Morris method and average marginal effects are compared. Comparison criteria are performance, available R packages or easy interpretability of results. Local methods concern one specific observation. LIME and Shapley values have been selected as local methods for analysis in this paper. The winning global and local method were then implemented and visualized in a dashboard, which can be found at https://juliafried.shinyapps.io/MunichRentIndex/.

In addition, the IML approach is compared with the model of the "original" Munich rent index 2017, which is based on simpler interpretable methods. This study shows that, model-agnostic methods provide explanations for machine learning models and the Munich rent index can be estimated with the IML approach. Model-agnostic interpretable machine learning offers enormous advantages because the underlying models are interchangeable and complex patterns in data can be explained globally and locally.

Keywords: *Interpretable Machine Learning; Black Box Models; Munich Rent Index; Shapley value.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

Enhancing UX of analytics products with AI technology

Christo Mirchev, Jean Metz, Markus Herrmann
Global Data Science, GfK, Germany.

Abstract

Insights and knowledge extraction from conventional analytics or reporting solutions is mostly neither trivial, nor intuitive. Moreover, many applications have unique interfaces and operating controls, forcing users to understand the tool's domain language and handling procedures, in order to find specific information. Such complicated handling creates cognitive load and impacts the users' productivity.

More specifically due to the complexity of the purpose of analytics products, to provide meaningful information (e.g. descriptives, predictions, prescriptions) at the right time, it must be considered that users' journeys in analytics products fundamentally differ to the journeys of users of traditional e-commerce products. Whereas a common rule- or filtering based recommendation routine, or a chatbot, might be applicable to facilitate and enhance the overall User Experience (UX) of online shoppers, this might not suffice for analysts who are seeking to derive insights from data.

We present preliminary results of an industry research study about the approach to combine natural language dialog- and content-flow based user interactions with content recommendations, in order to enhance UX of information retrieval from a data-driven analytics system. We demonstrate a prototype model towards a virtual assistant system that integrates predictions of the user's intention which information to retrieve next with prescriptive analytics based on the context of the current and past conversations.

Keywords: *Machine Learning Engineering; UX; Conversational AI; Natural Language Understanding.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

Search in second Hand market : The case of mobile phone

Yeon Ju Baik
Department of Economics, University of Wisconsin-Madison, United States.

Abstract

In this paper, I analyzed the search behavior of used mobile phone sellers and buyers in online trading platform by using sequential search model. The identification of search is achieved by variations in the duration and posted prices. I find that sellers face different selling costs which are mainly induced by the competition in the market. When competition level among the sellers is high, the more price dispersion is observed. It is consistent with the literature that studied positive relationship between the number of sellers and price dispersion. In equilibrium, consumers have higher search costs for the less competitive brand.

Keywords: *smartphone; secondhand; search.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

The epistemological impacts of big data on public opinion studies

Pedro Caldas, Vinicius Romanini
School of Communication and Arts, University of São Paulo, Brazil.

Abstract

In this work, we seek to highlight and describe the main differences between traditional public opinion polls (made by using methods and techniques traditionally undertaken in the social sciences), and those accomplished through methodological processes made possible by the adoption of big data. We ensure a special focus on the consequences brought about by the use of nonparametric analysis over parametric analysis to show how big data is impacting not only the methodological aspects but the epistemological basis of public opinion studies in general. Researchers see an epistemological struggle between methodology and theory in public opinion studies. This struggle is composed of two approaches: a quantitative one and a qualitative one. On the one hand, we have quantitative polls methods which lead to an excessively contextual representation of public opinion. On the other hand, we have general theories that do grasp public opinion in most of its complexity but fall short in providing sophisticated empirical tools for contextual analysis of public opinion specific issues. The methods undertook by pollsters, as many others used in social sciences rely upon classical scientific structures, where researchers conduct their studies through hierarchical theories and survey techniques to access and understand their subject. In these cases, the researchers must pose the research problem a prioristically, to parametrize and create the questionnaires before the collecting of the data to be analyzed after. By using big data models, the need for posing a research problem and parametrize the proceedings of the study a prioristically no longer exists, thus contributing to a characterization of public opinion that is qualitative and way more complex, rather than the traditional one. Although not yet strictly statistically representative, public opinion studies made by using datasets collected from social media provide us with a view of public opinion that shows, among other things, the main actors (persons, groups, and organizations), their powers of influence over the others and their interests in public opinion formation movement.

Keywords: *Public opinion; big data; polls; mining; epistemological impacts*.

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

Measuring and Forecasting Job-Search in Italy using Machine Learning

Carlo Drago[1, 2], Gentian Hoxhalli[1, 3]
[1]University "Niccolò Cusano", Rome Italy, [2]NCI University London, United Kingdom, [3]Luarasi University, Tirana, Albania.

Abstract

The Social Media are becoming more and more important to allow to measure phenomena which are very difficult to measure on a different way. In this sense these data can become relevant indicators which could be used in the analysis of the business cycle. In this work we will consider data related job-search queries on Google, in order to measure the intensity of the job-search behavior over the time. From the queries we are able to identify how vary during the business cycle the job-search behavior using Google. These behaviors are very relevant because they can lead to changes in job positions (transition from unemployed to the employed but also transitions on the job employed-employed). So forecasting this behavior we can have tools to interpret and analyze the business cycle. Finally we consider a forecasting approach based on Machine Learning in order to predict over the time the job-search behavior. The different forecasting approach considered are compared and finally validated by considering the forecast adequacy of the different predictions obtained.

Keywords: *Job-search; labor market; business-cycle; forecasting; machine-learning; forecast adequacy.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València