CARMA 2023
Sevilla

Proceedings of the 5th International Conference
on Advanced Research Methods
and Analytics

edUPV
Universitat Politècnica de València

28 June – 30 June, 2023
Sevilla, Spain

UNIVERSIDAD DE SEVILLA

Congress UPV

5th International Conference on Advanced Research Methods and Analytics (CARMA 2023)

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

Rocío Martínez-Torres
Sergio Toral

Publisher

2023, Editorial Universitat Politècnica de València
www.lalibreria.upv.es / Ref.: 6385_01_01_01

ISBN: 978-84-1396- 086-9
ISSN: 2951-9748

DOI: http://dx.doi.org/10.4995/CARMA2023.2023.17009

Preface

Rocío Martínez-Torres [1], Sergio Toral [2]
[1] Dept. of Business Administration and Marketing, University of Seville, Spain [2] Dept. of Electronic Engineering, University of Seville, Spain

Abstract

The development of the Internet and Big Data information sources is driving new and increasingly interdisciplinary research methods in economics and the social sciences. The 5th 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.

Keywords: *Big Data sources, Social Media mining, Natural Language Processing, Digital transition and global society, google trends, Big Data applications.*

1. Preface to CARMA2023

This volume contains the selected papers of the Fifth International Conference on Advanced Research Methods and Analytics (CARMA 2023) hosted by the University of Seville, Spain during 28, 29 and 30 June 2023. This fifth 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 after the digital transformation.

The selection of the scientific program was directed by Sergio Toral, who led an international team of 54 scientific committee members representing institutions worldwide. Following the call for papers, the conference received 76 paper submissions from all around the globe. All submissions were reviewed by the scientific committee members under a double-blind review process. Finally, 57 papers were accepted for oral presentation during the conference, ensuring a high-quality scientific program. It covers a wide range of research topics on the Internet and Big Data, including digital transition and global society, google trends, travel, tourism and leisure, social media, public opinion mining, Natural Language Processing, among others. Additionally, 8 papers with promising work-in-progress research were selected for presentation during the conference.

The scientific program included two keynote speakers that reviewed the state-of-the-art techniques and applications of the Internet and Big Data. The first keynote was Anastasija Nikiforova (University of Tartu, Estonia) who overviewed the latest public data ecosystems in and for smart cities. The second keynote speech was delivered by Patrick Mikalef (Norwegian Universtiy of Science and Technology, Norway) and dealed with Responsible Artificial Intelligence and Big Data Analytics.

CARMA 2023 also featured a special session on “How can innovative data collection and analysis methods support evidence-based policymaking in the EU?” chaired by Paolo Canfora, from the Joint Research Centre, European Commission. This session offered a complementary institutional perspective on how to use the Internet and Big Data sources and methods for public policy.

The conference organizing committee would like to thank all who made this fifth edition of CARMA a great success. Specifically, thanks are indebted to the authors, scientific committee members, special session organizer, invited speakers, session chairs, reviewers, presenters, sponsors, supporters, and all the attendees. Our final words of gratitude must go to the Faculty of Economics and Business of the Universidad de Sevilla for supporting CARMA 2023.

2. Organizing Committee

General Chair

M. Rocío Martínez Torres, Universidad de Sevilla

Scientific Committee Chair

Sergio L. Toral Marín, Universidad de Sevilla

Local Organization

Pedro Baena Luna
Lourdes Cauzo Bottala
Myriam L. González Limón
F. Javier Quirós Tomás
Samuel Yanes Luis

3. Steering Committee

Aidan Condron
Caterina Liberati
Giuliano Resce
Josep Domenech
Juri Marcucci
Lisa Crosato
María Olmedilla Fernández
María Rosalía Vicente
Markus Hermann
Pablo de Pedraza García
Rocío Martínez Torres
Sergio Toral Marín

4. Sponsors and Supporters

University of Seville
Facultad de Ciencias Económicas y Empresariales, Universidad de Sevilla
Instituto de Economía y Negocios de la Universidad de Sevilla (IUSEN)
Departamento de Administración de Empresas y Marketing, Universidad de Sevilla
Universitat Politècnica de València
Joint Research Centre, European Commission

Cátedra Metropol Parasol
Cátedra de Digitalización Empresarial IBM
Lidl
Coca-Cola

4. Scientific Committee

Agustín Indaco, Carnegie Mellon University, Qatar
Ana Debón Aucejo, Universitat Politècnica de València
Ana Isabel Irimia-Dieguez
Anna Rosso, Università degli Studi dell'Insubria
Anto Aasa, University of Tartu
Antonio de Nicola, ENEA, Italia
Antonino Virgillito, Agenzia delle Entrate - Italian Revenue Agency
Benito Mignacca, University of Cassino and Southern Lazio
Carlo Drago, University "Niccolò Cusano", Rome
Caterina Liberati, University of Milano-Bicocca
Catherine Beaudry, Polytechnique Montreal
Dharma Vázquez Torres, University of Puerto Rico
Enrique Orduña-Malea, Universitat Politècnica de València
Esther Chavez Miranda, University of Seville
Esther García del Río, University of Seville
Federico Neri, Deloitte, Italy
Féliz Velicia Martín, University of Seville
Fernando Almeida, University of Porto & INESC TEC
Francisca García-Cobián Richter, Case Western Reserve University
Francisco Javier Rondán Cataluña, University of Seville
Francisco Javier Quirós Tomás, University of Seville
Gabriel Cepeda Carrión, University of Seville
Giuliano Resce, Italy
Hugo Álvarez-Pérez, Egade Business School - Tecnologico de Monterrey
José A. Álvarez-Jareño, University of Valencia
Jose Carlos Romero Moreno, University of Paris-Dauphine-PSL,
José Luis Ortega, CSIC
Josep Domènech, Universitat Politècnica de València
Juan Fernández de Guevara, Universitat de Valencia & Ivie
Juri Marcucci, Bank of Italy
Konstantinos Tsagarakis, Technical University of Crete
Lisa Crosato, Ca' Foscari University of Venice

Index

Digital transition and global society

Big Data applications (I)

Google Trends

Big Data sources

Natural Language Processing (I)

Advances in travel, tourism and leisure

Big Data Sources: Social Media

Big Data Applications in Education

Natural Language Processing (II)

Big Data sources: public opinion mining

Big Data Sources in Economics and Social Sciences

Big Data Methods. Machine Learning

Big Data Applications in Prices

Big Data Applications (II)

Big Data Methods

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16496

An energy transition without externalization?

Alejandro Caballeros-Finkelstein[1], Jesús Ramos-Martín[2,3], Cristina Madrid-López[1*]
[1]LivenLab (PID2020-119565RJ-I00), SosteniPra group (2021SGR00734), Institut de Ciència i Tecnologia Ambientals (ICTA-UAB)(CEX2019-0940-M), Universitat Autònoma de Barcelona, Spain. [2]ICTA-UAB. [3]Department of Economics and Economic History, Universitat Autònoma de Barcelona, Spain.

*Correspondence to: cristina.madrid@uab.cat

Abstract

The extended abstract presents the first results of the estimation of externalization levels of Spain through its energy transition towards 2030. . It represents a first step towards building the Calliope Spain model adapted from EuroCalliope. This first portion of the work presents the trade emissions balance for Spain for 2010 and 2015 using an Input-Output methodology. The results position Spain as a "net exporter" of emissions given that the country imports more goods and services than it exports. These insights serve as baselines to establish the country´s total internal and externalized emissions by trade partner as will be used in Calliope Spain. By examining these trends, it is possible to gain valuable insights into the correlation between international trade and greenhouse gas emissions in Spain's energy sector. This first analysis concludes with recommendations for policy.

Keywords: *Multi.Regional Input-Output; Calliope Energy Model; Trade emissions balance; Big Data*

1. Introduction

As part of its energy transition strategy, the EU has set that by 2030 member states should reduce their greenhouse gas emissions by 45% of their 1990 levels, and zero by 2050 (European Council, 2022). Policymakers rely on energy models to find what configurations of the energy system will help us reach this target. Energy system optimization models (ESOMs) process a big amount of data to reach an optimum energy system pathway according to a number of constraints. As of today, most of the ESOMs do not include environmental issues nor international trade in proper detail. One of these models is Calliope, an open-source, transparent framework for the modelling of energy systems, which also does not include environmental issues and international trade beyond the need for imports of electricity and fuel and GHG emissions.

International trade plays a double role in the energy transition. On one hand, it has traditionally increased greenhouse gas emissions through production, (often long route) transportation, and consumption of imported and exported goods and services. On the other hand, it has limited the local impacts of activities, by "externalizing" environmental impact to countries with fewer environmental regulations or lower labor standards. Thus the lack of inclusion of external trade in models that guide the energy transition seems to result in an incomplete picture of the impacts of future energy pathways.

Spain has a significant impact on global greenhouse gas emissions through the consumption of imported products. Current estimates account for a 160% economic trade deficit growth in 2022 (Ministerio de Industria Comercio y Turismo, 2023) and in 2018 imported goods were responsible for 50% of the country´s total greenhouse gas emissions (Ministerio para la Transición Ecológica y Reto Demográfico, 2020).

Spain is set to curve its emissions through the implementation of The Integrated National Energy and Climate Plan (PNIEC), which proposes the reduction of greenhouse gas emissions by 23% compared to 1990 levels by 2030 and zero by 2050 (Ministerio para la Transición Ecológica el Reto Demográfico, 2020). However, these targets refer to domestic emissions only. Neither the PNIEC nor the TIMES model used for its definition includes the externalization of environmental impacts elsewhere. Still, following the Sustainable Development Goals (SDGs), energy systems must be designed in a way that are both clean and fair. Externalization of impacts is one of the most common forms of modern colonialism (Muradian & Martinez-Alier, 2001) and must be considered in the definition of cleaner (and fairer) energy systems.

In this work, we take the energy transition scenarios in Spain for 2030 as modelled by Euro Calliope (Pickering et al., 2022) and calculate their level of externalization assuming the productive structure and terms of trade and compare it with a previous study for the year 2010 and 2015. This analysis is a first step towards building Calliope Spain, an adaptation

of EuroCalliope model that will include high-level resolution externalization. Analyzing these trends, offer insights into the relationship between international trade and greenhouse gas emissions in Spain for energy and concludes with policy recommendations.

2. Literature Review

The relationship between greenhouse gas emissions and international trade has been the subject of numerous studies and academic research in recent years. Several studies have highlighted the role of international trade in driving greenhouse gas emissions. For example, researchers have found that international trade is responsible for a significant portion of global greenhouse gas emissions, with emissions embodied in trade accounting for over a quarter of global emissions (Hertwich & Peters, 2009). Another study found that emissions embodied in China's exports were responsible for 26% of the country's total carbon dioxide emissions (Peters et al., 2011).

The estimation of greenhouse gas emissions in international trade has been extensively studied, with a particular interest in accounting for embodied emissions in trade - the emissions generated in the production and transport of goods produced in one country but consumed in others. Studies have looked at the amount of carbon embodied in international trade flows from different countries to determine the magnitude of emissions from imports (Peters et al., 2012; Wyckoff & Roop, 1994). The trade emissions balance or pollution terms of trade help to understand the environmental costs that shift between countries in international trade (Duan & Jiang, 2017). The diverse methodologies on the estimation of the trade emissions balance give research a wide variety of indicators and information to choose from, including revisiting the Leontief Input-Output model (Muradian et al., 2002; Muradian & O'Connor, 2001; Sánchez-Chóliz & Duarte, 2004).

Overall, the relationship between greenhouse gas emissions and international trade is a complex and multifaceted issue, with various factors influencing the extent of the relationship. Using the Euro Calliope model, it is possible to analyze different scenarios for the Spanish energy system, such as the impact of increasing renewable energy capacity, changing electricity demand patterns, or implementing different policy measures (Pickering et al., 2022). For example, the model can be used to evaluate the potential for increasing wind and solar capacity in Spain, and to explore the optimal mix of different energy sources to meet electricity demand while minimizing costs and greenhouse gas emissions.

3. Methodology

We revisit Serrano and Dietzenbacher´s trade emissions balance methodology for the small country case (Serrano & Dietzenbacher, 2010). The case considers a world economy

consisting of two regions ($r, s = 1,2$) where the country of interest is region 1 and the Rest of the World (RoW) is region 2. Each region is composed of n sectors that produce one product that might be used by other sectors as intermediate input or consumed (either at home or abroad). The model[1] $x = Ax + y$ for the estimation of the trade emissions balance of region 1 with region 2 is solved by $x = (I - A)^{-1}y$, where $L = (I - A)^{-1}$ is the Leontief inverse. The estimation associated with the production of each region is rendered from the multiplication of the gross output with a matrix of atmospheric emission coefficients defined as W^r. Each element of the matrix indicates the domestic emission of a pollutant per unit of an industry´s output for one of the regions. An important assumption is that the production technology and emission intensities are the same for the country and the RoW. The assumption is made due to a lack of data on technology or the RoW. The other assumption for this case is that the small country´s exports are considered negligible when compared to the RoW. Under these assumptions we formulate the trade emissions balance, and using Serrano and Dietzenbacher´s simplified expression we have:

$$1) \quad eb^1 = W(I - A - M)^{-1}(exp^1 - imp^1)^2$$

The variable exp^1 gives the vector of total exports and imp^1 the vector of total imports of the country. The equation will result in the trade emissions balance per type of pollutant embodied in aggregate exports and imports of region 1.

The data sources of this paper are the 2010 and 2015 Input-Output Tables from the Spanish National Statistics Institute as well as the Accounts of emissions into the atmosphere by branches of activity (Instituto Nacional de Estadística, 2015, 2018, 2022). The Input-Output Tables are categorized into 64 sectors (NACE) and 64 products (CPA), and the emissions accounts are categorized into 63 sectors and 13 pollutant substances. The estimations in this paper will consider 9 of the 13 pollutant substances (i.e., CO_2, CH_4, N_2O, SF_6, HFCs, PFCs, SO_2, NO_x, and NH_3) and will not consider sector 64 named "Activities of extraterritorial organizations and bodies" due to lack of data in the matrices. Taking this into account, the estimation is set by a 63 × 63 symmetrical environmental input-output table. In the following section, the empirical results will be discussed.

[1] We define x as the gross output of a country, A as the matrix of input coefficients and y as the final demand.

[2] For the simplified expression, M represents the import coefficients of the country. Adding A to M results in the technical input matrix that is then used in the Leontief inverse.

4. Provisional results

Using the information from the Input-Output Tables we aggregate all the NACE sectors' exports and imports for 2010 and 2015 to obtain the net exports for both years. We notice that in general imports are larger than exports in Spain for both years, rendering a negative trade balance. The data also shows two important shifts between both years. Firstly, both exports and imports increase between 2010 and 2015. Secondly, the increase in exports is greater than the imports, which then results in a decrease in the net exports between both years. We also consider the aggregate emissions per type of pollutant using the Accounts of emissions. By adding the total emissions per pollutant per sector, we compute the total emissions per type of pollutant (expressed in thousand tons), as seen in columns (1) and (2) of Table 1. The "Aggregate emissions per type of pollutant" show that the largest proportion of emissions is CH_4 for both years, notwithstanding the significant interannual decrease. Also, most of the emissions per type of pollutant decrease, except for SO_2 and NH_3. These changes are non-trivial because they affect the results in the trade emissions balance. In columns (3) to (8) we show the results from using Equation 1) and the selected datasets for 2010 and 2015:

Table 1. Aggregate emissions per type of pollutant and Trade emission balances (2010-2015). (Thousand tonnes)

	Aggregate emissions per type of pollutant		Emissions embodied in exports		Emissions embodied in imports		Trade emission balance	
	2010	2015	2010	2015	2010	2015	2010	2015
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CO_2	217	214	147,181,607	197,265,972	169,558,395	199,385,248	-22,376,789	-2,119,277
CH_4	1,537	1,492	678,027,460	883,919,393	781,111,517	893,415,556	-103,084,057	-9,496,163
N_2O	55	55	147,181,607	190,766,377	169,558,395	192,815,827	-22,376,789	-2,049,450
PFC	0.10	0.09	80,046	78,809	92,216	79,656	-12,170	-847
HFC	15	8	8,842,934	5,891,324	10,187,372	5,954,616	-1,344,438	-63,292
SF_6	0.23	0.22	251,621	306,624	289,876	309,918	-38,255	-3,294
NO_x	880	792	484,296,687	609,524,996	557,926,842	616,073,273	-73,630,155	-6,548,277
SO_2	249	262	172,857,204	271,909,577	199,137,588	274,830,768	-26,280,384	-2,921,191
NH_3	424	443	174,685,168	254,287,229	201,243,466	257,019,100	-26,558,299	-2,731,870

Source: Own elaboration from the 2010 and 2015 Spanish Input-Output and Accounts of emissions.

Columns (3) to (8) present the aggregate emission balances for Spain in 2010 and 2015. The data shows that, for both years, the emissions embodied in the imports are higher than those embodied in the exports. Spain has a negative trade emission balance for all gases in both years. This implies that Spain would be a "net exporter" of pollution to the RoW under the assumption that both regions use the same technology to produce commodities. The

goods that are produced in the exporting region create emissions within its geographical limits and are accounted for as such. Spain shifts the environmental cost of producing domestic final or intermediate consumption goods to other countries.

The imports of goods from Spain may contribute to greater carbon emissions in the RoW due to several factors, including transportation emissions and consumer trends. One of the primary contributors to carbon emissions in international trade is transportation. Goods from Spain need to be shipped to their destination, and the emissions from this transportation can be significant. The carbon footprint of transporting goods over long distances depends on the mode of transport, the distance traveled, and the weight and volume of the goods. For example, shipping by sea is less carbon-intensive than air transportation, but it can still result in significant emissions over long distances. Also, the import of goods from Spain may lead to increased consumption and associated carbon emissions. When consumers have access to a wider range of products, they are more likely to purchase them, which can increase the demand for transportation and production. This increased demand can lead to an increase in carbon emissions, particularly if the goods are produced in a carbon-intensive way.

Similarly, the use of imported inputs to export goods generates indirect emissions. These are considered in Equation 1) following Serrano and Dietzenbacher´s analysis (Serrano & Dietzenbacher, 2010). Given that the imports a greater proportion of goods than it exports, the emissions associated with them are also greater. The production of goods often involves the use of other goods and services that may themselves generate carbon emissions. For example, the production of steel for a car may involve the use of energy and raw materials that generate carbon emissions. The emissions associated with these upstream activities are known as indirect emissions, and they can be significant, particularly for goods that require a lot of resources to produce.

The results have a similar trend to those found in Serrano and Dietzenbacher. These cannot be compared directly due to the change in methodologies in the Input-Output Tables in 2008 (Instituto Nacional de Estadística, 2011). In general, Spain is a net importer country that renders larger “exports” of emissions to other countries that produced its imported goods. However, the analysis can be expanded to the emissions created by each sector. Serrano and Dietzenbacher criticize the limits of Equation 1), which holds for aggregate and broader estimations of the emissions of the economy´s trade emissions balance. To create more detailed estimations other methodologies can be adopted (Peters et al., 2012; Sánchez-Chóliz & Duarte, 2004; Serrano & Dietzenbacher, 2010).

5. Conclusions

The results of any modeling exercise are highly dependent on the assumptions and data inputs used and should be interpreted with caution and this is the case for energy modeling. In this work, we contribute to close a dangerous gap that excludes externalization in energy system optimization models. We revisited trade emissions balance estimation methodologies for the most recent data available for Spain. The results present the country as a net exporter of emissions through its import channels but not directly attributed to the consumption of imported goods. To reduce the carbon footprint of international trade, it is essential to account for emissions generated directly and indirectly. Thus the design of Calliope Spain must consider international trade in its design with enough level of detail to offer useful insights for the implementation of emission reduction policies.

Acknowledgments

This research has been funded by the Agencia Estatal de Investigación, Gobierno de España (TED2021-132032A-I00 and PID2020-119565RJ-I00 grants)

References

Duan, Y., & Jiang, X. (2017). Temporal Change of China's Pollution Terms of Trade and its Determinants. Ecological Economics, 132, 31–44. https://doi.org/10.1016/j.ecolecon.2016.10.001

European Council. (2022, December 6). European Green Deal. European Green Deal. https://www.consilium.europa.eu/en/policies/green-deal/

Hertwich, E. G., & Peters, G. P. (2009). Carbon footprint of nations: A global, trade-linked analysis. Environmental Science and Technology, 43(16), 6414–6420. https://doi.org/10.1021/es803496a

Instituto Nacional de Estadística. (2011). Contabilidad Nacional de España Base 2008 Características metodológicas.

Instituto Nacional de Estadística. (2015). Tablas Input-Output 2010.

Instituto Nacional de Estadística. (2018). Tablas Input-Output 2015.

Instituto Nacional de Estadística. (2022). Cuentas de emisiones a la atmósfera por ramas de actividad (CNAE 2009) y Hogares como consumidores finales, sustancias contaminantes y periodo.

Ministerio de Industria Comercio y Turismo. (2023). A MINISTERIO Deputy Directorate-General for Studies and Trade Policy Evaluation Report prepared by Secretariat of State for Trade EXECUTIVE DIRECTION. https://comercio.gob.es/ImportacionExportacion/Informes_Estadisticas/Paginas/Informes-periodicos.aspx

Ministerio para la Transición Ecológica el Reto Demográfico. (2020). Resumen del Estudio Ambiental Estratégico del Plan Nacional Integrado de Energía Clima.

Minsterio para la Transición Ecológica y Reto Demográfico. (2020). Inventario Nacional de Emisiones a la atmósfera. Emisiones de gases de efecto invernadero. Serie 1990-2018. Informe resumen.

Muradian, R., & Martinez-Alier, J. (2001). Trade and the environment: from a "Southern" perspective. In Ecological Economics (Vol. 36). www.elsevier.com/locate/ecolecon

Muradian, R., & O'Connor, M. (2001). Inter-country environmental load displacement and adjusted national sustainability indicators: concepts and their policy applications. International Journal of Sustainable Development, 4(3), 321–347. https://doi.org/10.1504/IJSD.2001.004445

Muradian, R., O'connor, M., & Martinez-Alier, J. (2002). Embodied pollution in trade: estimating the "environmental load displacement" of industrialised countries. In Ecological Economics (Vol. 41). www.elsevier.com/locate/ecolecon

Peters, G. P., Davis, S. J., & Andrew, R. (2012). A synthesis of carbon in international trade. Biogeosciences, 9(8), 3247–3276. https://doi.org/10.5194/bg-9-3247-2012

Peters, G. P., Minx, J. C., Weber, C. L., & Edenhofer, O. (2011). Growth in emission transfers via international trade from 1990 to 2008. Proceedings of the National Academy of Sciences of the United States of America, 108(21), 8903–8908. https://doi.org/10.1073/pnas.1006388108

Pickering, B., Lombardi, F., & Pfenninger, S. (2022). Diversity of options to eliminate fossil fuels and reach carbon neutrality across the entire European energy system. Joule, 6(6), 1253–1276. https://doi.org/10.1016/j.joule.2022.05.009

Sánchez-Chóliz, J., & Duarte, R. (2004). CO2 emissions embodied in international trade: Evidence for Spain. Energy Policy, 32(18), 1999–2005. https://doi.org/10.1016/S0301-4215(03)00199-X

Serrano, M., & Dietzenbacher, E. (2010). Responsibility and trade emission balances: An evaluation of approaches. Ecological Economics, 69(11), 2224–2232. https://doi.org/10.1016/j.ecolecon.2010.06.008

Wyckoff, A. W., & Roop, J. M. (1994). The embodiment of carbon in imports of manufactured products Implications for international agreements on greenhouse gas emissions.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16445

Analysis of challenges of digital service enabled by big data analytics technologies using a new integrated multiple-criteria decision-making (MCDM) method

Sara Saberi, Abbas Mardani
Business School, Worcester Polytechnic Institute, Worcester, MA 01609, USA.

Abstract

The digitalization of services and products is an approach adopted by modern companies to produce value. The key to success is knowing what your customers are saying about your company by compiling data in many aspects and reviewing the digital content collected from digitally enabled services. On the other hand, text review is a highly subjective task. The raw data has complex features, making analyzing the data on digital services a very complex and intriguing problem. This study collects the main challenges of digitally enabled services to offer an inclusive framework and describes the framework's potential in dealing with application-specific challenges. This study aims to suggest a data-driven decision-making model using the "intuitionistic fuzzy sets (IFSs)", "method based on the removal effects of criteria (MEREC)", "rank sum (RS), and the "multi-attribute multi-objective optimization with ratio analysis (MULTIMOORA)" approaches. The IF-MEREC-RS tool computes the weights of the digital service challenges that big data analytics technologies enable and the IF-MULTIMOORA method prioritizes the technologies to assess the challenges. Then, an integrated decision-making framework is developed to investigate these challenges' subjective and objective weights using expert opinion. Using big data analytics, the proposed model can assess the preferences of technologies over different challenges.

Keywords: *Digital service; big data analytics; social media; digital technologies; data-driven decision-making.*

1. Introduction

In recent years, every facet of business and organizational activities has been digitized, which has resulted in the creation of huge datasets analysis purposes. Through big data and analytical procedures, these datasets can provide insights for offering sustainable value to enhance business performance and competitive benefits (Wamba et al., 2017). The recent literature is enriched with "big data analytics (BDA)" because of the massive acceptance of the Internet as well as the emergence of Web 2.0 technologies. Both academicians and practitioners are greatly interested in BDA due to the increased demand for understanding the trends in massive datasets (Ghani et al., 2019). More and more data are being compiled in many domains, such as supply chains, health care, and finance, due to the new developments in cyber-physical systems, sensing networks, and IoT. Though, the data gathered this way suffer from an inherent uncertainty because of incompleteness, noise, and inconsistency. To effectively analyze such data, there is a need for progressive analytical approaches to efficiently review and/or predict future courses of action with high accuracy and innovative decision-making policies. With a great and fast increase in the amount and variety of data and, consequently, the increase of its uncertainty degree, the outcomes of analyses and also the decisions made accordingly lack confidence. It is not easy to conduct big data analyses with the use of conventional data analytics (Tsai et al., 2015). This failure is due to the fact that the conventional methods may lose effectiveness because of the five V's characteristics of big data, i.e., "high volume, low veracity, high velocity, great variety, and high value" (Chen et al., 2014; Ma et al., 2014).

Big data also has other features, e.g., viability, validity, viscosity, and variability (Djafri & Gafour, 2022; Xin et al., 2021). A number of "artificial intelligence (AI)"-based techniques, for instance, data mining, "natural language processing (NLP)", "machine learning (ML)", and "computational intelligence (CI)", have been developed to offer BDA solutions due to their higher speed, accuracy, and precision when applied to massive data (Chen et al., 2014). These techniques typically aim to discover the information, indefinite correlations, and hidden patterns in massive datasets (Tsai et al., 2015).

Digital service has a leading role in our daily lives, leading to massive data generation. The big data associated with digital service finds the most progressive applications in the socio-economic domains. Many studies have been carried out on the challenges that inherently exist in specific applications of digital service or big data separately; however, the literature lacks research into digital service enabled by big data analytics. To bridge this gap, the current study discusses the latest digital service enabled challenges by big data analytics applications used in the industry 4.0 era. In addition, this study presents an inclusive framework of digital service enabled by big data analytics technologies and describes its potential to deal with application-specific challenges. This study aims to suggest a data-driven decision-making model for the evaluation of the multi-attribute decision analysis (MADA) problem. In this

line, the proposed method will discuss computing the weights of the digital service-enabled challenges by big data analytics technologies. Then, the proposed method prioritizes the technologies to assess the digital service challenges that big data analytics technologies enable. In this regard, this study aims to suggest an integrated framework based on the removal effects of criteria (MEREC)", and the "multi-attribute multi-objective optimization with ratio analysis (MULTIMOORA)" approaches called the "MEREC-MULTIMOORA" for the evaluation of the MADA problem. In this line, the MEREC tool is discussed to compute the weights of challenges of SM in the era of BDA technologies.

2. Literature review

2.1. Digital service enabled by big data analytics

Big data refers to huge or complicated data sets that typically surpass conventional systems' technical capability in the storage, processing, management, interpretation, and visualization of data (Kaisler et al., 2013). At present, we face an exponentially growing trend in the volume of data, which is expected to reach zettabytes per year in a few years. Scholars and practitioners believe such an overflow of data brings about new opportunities; for that reason, numerous companies are attempting to improve their BDA capacities to understand the hidden values of big data better. Kambatla et al. (2014) comprehensively discussed the trends in BDA, including both software and hardware. Two challenging tasks are collecting and storing data from widely-distributed sources in the storage systems and running a diverse set of computations. Zhong et al. (2016) investigated the currently used big data technologies, including those introduced for storing, processing, and visualizing data. There is still a need to systematically review novel analytical methods, tools, and techniques to discern decisions in different domains (Hagel, 2015).

Big data and progressive techniques of data analytics could be applied to developing analytical and computational models (Iqbal et al., 2020). The literature shows that there is still interest in finding the best ways to develop the infrastructure, which has led to the introduction of different data mining and ML algorithms in various study areas.

Big data is mainly focused on the psychological aspect of predicting the consumers' requirements rather than understanding them. It is essential to investigate how the customers behavior and the things they will buy next after purchasing the goods could be predicted. This will help to understand the consumers' perception regarding the brand, and it will show the way to enhance the quality and effectiveness of target advertising.

A score was developed by Scholz et al. (2018) covering the position effect of digital service such as social media (SM) inside. This score helps to analyze the inside impact on individuals and firms. The results of the study by Goldberg (1990) give the social facilitation inspiration,

participating and socializing inspiration, and information inspiration that pressures the consumers' common attitudes in the direction of SM sites. It had a well-constructed outcome on their attitudes in the direction of marketers social networking sites. In SM posts, there are many potentials that could be used for data mining and analysis. With understanding such potentials, platform providers tend to put a limitation on individuals access to such data. This shift causes new challenges for social scientists and other non-profit scholars seeking to analyze public posts to better understand human interactions and improve human conditions. SM analytics is a research axis that is concentrated on extracting insights from SM-induced data to aid individuals and organizations in making the best decisions about different disciplines of life. There is a need for big data technologies to be applied. For that reason, the current study aims to help researchers working in this field discover the challenges faced with data analysis using big data technologies. A comprehensive review was conducted to collect the challenges and obstacles faced when integrating big data technologies with digitally enabled services, and the result is presented in Figure 1. To the best of our knowledge, this set of challenges is the main contribution of this study. However, we need to look for practical multi-criteria decision-making tools to examine the subjective and objective weights of these challenges. The conventional MCDM methods are proposed to aid decision-makers in making the best decisions in different situations. Though, most decision-making situations necessitate considering the decision experts (DE) experiences and judgment. The use of the Fuzzy sets theory could accomplish this.

3. Research method

Fuzzy sets (FSs) and their generalizations feature can help handle information that suffers from incompleteness and imprecision. However, not all fuzzy multi-criteria decision-making tools can be applied to incomplete and uncertain data, which may appear recurrently in real situations. To effectively address such challenges, a robust formal general framework was developed by Atanassov (1986) as a novel branch of mathematics, termed "intuitionistic fuzzy sets (IFSs)", to treat the problem with uncertainty and ambiguity of information. IFSs define each object with a "membership function (MF)", a "non-membership function (NF), and an "indeterminacy function (IF)" to reflect the unknown/neutral environment. Later on, an IFS MCDM model was introduced by Mishra et al. (2021) for ranking and assessing suppliers using the "combinative distance-based assessment (CODAS)" tool.

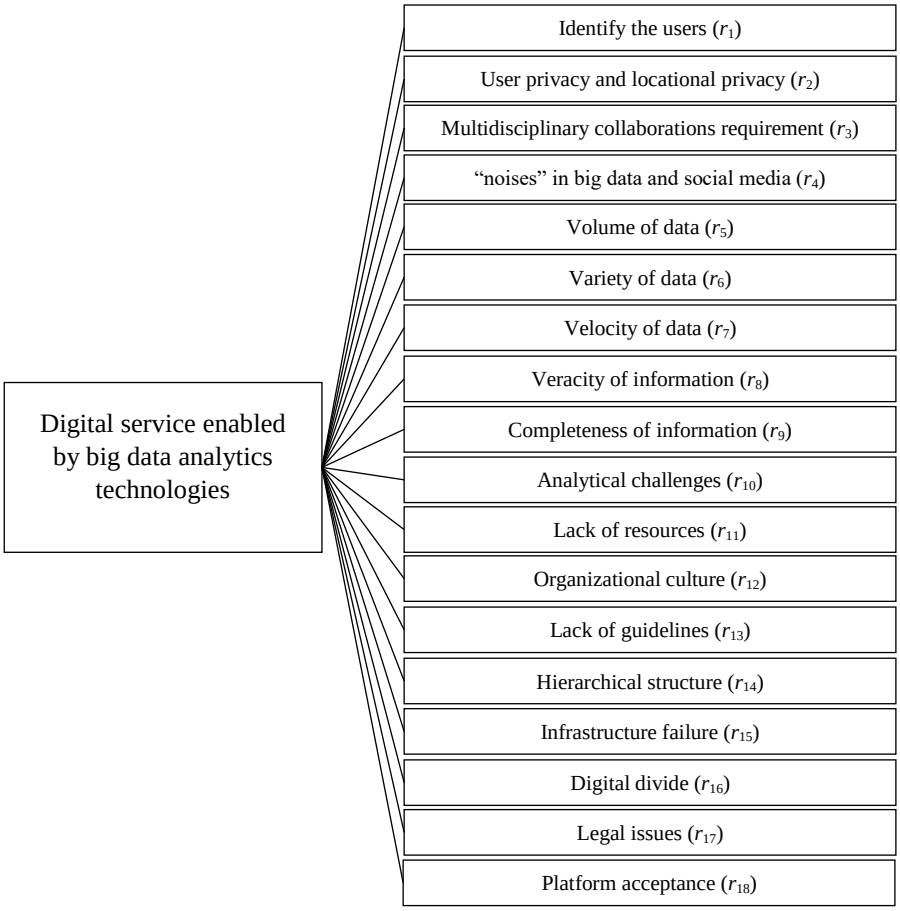


Figure 1: Challenges of digital service enabled by big data analytics technologies

Rani et al. (2021) also proposed the Fermatean fuzzy information-MEREC-additive ratio assessment (ARAS) method to provide a solution to the problem of selecting the best available method for the treatment of food waste. Mishra et al. (2022) developed an integrated MEREC-MULTIMOORA tool for the selection of "low-carbon tourism strategies" in the "single-valued neutrosophic sets (SVNSs)" setting. In this proposed MCDM method, the MEREC is utilized for finding the objective weighting coefficients. The ranking sum (RS) method is employed to find the subjective weighting coefficient.

The RS developed by Stillwell et al. (1981) is used for ranking values of selected criteria with the help of the decision maker's opinions. Later, Hezam et al. (2022) introduced a hybrid MCDM methodology by combining the "MEREC-RS-DNMA" approach with IFSs and applied it to evaluate the "alternative fuel vehicles (AFVs)" problem.

In 2006, the "multi-objective optimization with the ratio analysis (MOORA)" tool by combing the "reference point (RP)" and "ratio system (RS)" was introduced in W. Brauers and Zavadskas (2006) study to treat the MCDM issues. W. K. M. Brauers and Zavadskas (2010) renamed MOORA into "multiattribute multi-objective optimization with the ratio analysis (MULTIMOORA)" by adding "full multiplicative form (FMF)". (W. K. M. Brauers and Zavadskas (2010) validated that the MULTIMOORA has higher efficiency, and better stability with advanced robustness by comparing it with other tools such as the "technique for order of prioritization by similarity to ideal solution (TOPSIS)", the "analytic hierarchy process (AHP)", and the "VIKOR (visekriterijumska optimizacija I kompromisno resenje in Serbian)".

3.1. Proposed data-driven decision-making model

We are working on a new methodology to propose MEREC-MULTIMOORA, an extended decision-making methodology. MULTIMOORA properly integrates the benefits of various aggregation functions. The final integration function of MULTIMOORA extensively considers the utility values and the alternative ranks; this will cause the final ranking result highly reliable considering the attribute of data collected from digitally enabled services.

3.2. Case study

There are several challenges for SM in the era of BDA technologies; therefore, this study will implement a survey approach with the current literature review and interviews with experts to identify these challenges. In the first step, we discuss the important challenges for SM in the era of BDA technologies using the current literature review. In the following stage, we will send the identified challenges to different experts to select the most important challenges for SM in the era of the BDA technologies section. A total of 25 analytic data managers from different online learning websites will be invited. In the next stage, we will invite four DEs in the area of SM and BDA to evaluate the identified challenges.

4. Conclusion

The biggest proportion of big data surge is in words, videos, images, or a combination of them. Big data is exponentially growing, and it is expected even to accelerate its growth pace in the future. Only a small amount of digital service-generated data is analyzed effectively; however, business managers believe that these data greatly support making intelligent decisions if correclty analyzed. Several tools are being designed and proposed in the literature for using digital service-produced data to gauge consumers behavior and turn it into actionable information. For the analysis, ranking, and evaluation of the most important challenges that arise in digital service in the BDA technologies era, the current paper proposes an integrated decision-making method to analysis the challenges of different online

learning websites. Big data technology helps organizations to take deeper insights from consumer feedback. Accordingly, a decision-making model will be introduced using the MEREC and the MULTIMOORA tools called the MEREC-MULTIMOORA method to evaluate the main challenges of digital service in the era of BDA technologies in different online learning websites. To rank the main challenges of digital service in the era of BDA technologies online learning websites, the MEREC is applied, and the MULTIMOORA method is used to find the rank of different technologies over different challenges.

References

Brauers, W., & Zavadskas, E. K. (2006). The MOORA method and its application to privatization in a transition economy. *Control and Cybernetics, 35*, 445-469.

Brauers, W. K. M., & Zavadskas, E. K. (2010). Project management by MULTIMOORA as an instrument for transition economies. *Technological and Economic Development of Economy*(1), 5-24.

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications, 19*(2), 171-209. doi:10.1007/s11036-013-0489-0

Djafri, L., & Gafour, Y. (2022, 2022//). *Machine Learning Algorithms for Big Data Mining Processing: A Review.* Paper presented at the Artificial Intelligence and Its Applications, Cham.

Ghani, N. A., Hamid, S., Targio Hashem, I. A., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior, 101*, 417-428. doi:https://doi.org/10.1016/j.chb.2018.08.039

Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. *J Pers Soc Psychol, 59*(6), 1216-1229. doi:10.1037//0022-3514.59.6.1216

Hagel, J. (2015). Bringing analytics to life. *Journal of Accountancy, 219*(2), 24-25. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994684186&partnerID=40&md5=a462167597775aedcf5cfb97551e14bc

Iqbal, R., Doctor, F., More, B., Mahmud, S., & Yousuf, U. (2020). Big Data analytics and Computational Intelligence for Cyber–Physical Systems: Recent trends and state of the art applications. *Future Generation Computer Systems, 105*, 766-778. doi:https://doi.org/10.1016/j.future.2017.10.021

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, 7-10 Jan. 2013). *Big Data: Issues and Challenges Moving Forward.* Paper presented at the 2013 46th Hawaii International Conference on System Sciences.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing, 74*(7), 2561-2573. doi:https://doi.org/10.1016/j.jpdc.2014.01.003

Ma, C., Zhang, H. H., & Wang, X. (2014). Machine learning for Big Data analytics in plants. *Trends in Plant Science, 19*(12), 798-808. doi:10.1016/j.tplants.2014.08.004

Scholz, M., Schnurbus, J., Haupt, H., Dorner, V., Landherr, A., & Probst, F. (2018). Dynamic effects of user- and marketer-generated content on consumer purchase behavior:

Modeling the hierarchical structure of social media websites. *Decision Support Systems, 113*, 43-55. doi:https://doi.org/10.1016/j.dss.2018.07.001

Tsai, C.-W., Lai, C.-F., Chao, H.-C., & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big Data, 2*(1), 21. doi:10.1186/s40537-015-0030-3

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J.-f., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research, 70*, 356-365. doi:https://doi.org/10.1016/j.jbusres.2016.08.009

Xin, Z., Xiaohong, L., & Chenming, G. (2021, 16-18 April 2021). *Research on Mobile Learning of College Students based on Wechat.* Paper presented at the 2021 International Conference on Internet, Education and Information Technology (IEIT).

Zhong, R. Y., Newman, S. T., Huang, G. Q., & Lan, S. (2016). Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Computers & Industrial Engineering, 101*, 572-591. doi:https://doi.org/10.1016/j.cie.2016.07.013

The digital divide: An approach through machine learning classifiers

Andrés Aleán[1], Manuel Vicente Nieto Mengotti[2]
[1]IDEEAS, Universidad Tecnológica de Bolívar, Colombia, [2]Department of Economics, Universidade da Coruña, Spain

Abstract

In 2022, 2.9 billion people worldwide lacked access to the internet, thus being unable to benefit from the digital economy (WEF, 2022). Moreover, lacking internet access at home can further exacerbate existing educational and economic inequalities. Thus, it is crucial not only to identify the sociodemographic profile of households that lack internet access, but of those most vulnerable to lacking internet access in the future (Hidalgo et al., 2020). This study applies several widely used machine learning classifiers (logit regression, naïve Bayes, linear discriminant analysis, k-nearest neighbors and random forest; James et al., 2021) to analyze the main socioeconomic internet access drivers for the Mexican population, using household surveys for the period between 2016 and 2020 (INEGI, 2020). Our principal result is that income, education level, and rurality are the main factors determining lack of internet access, both present and future; and that gender and occupation only play a secondary role in explaining the digital divide. These results can inform the formulation of public policies with the aim to secure universal access to the internet, and thus prevent the widening of existing inequalities in development.

Keywords: *Internet access; Machine learning; Forecasting and nowcasting.*

This document was prepared within the framework of Manuel Vicente Nieto Mengotti's postdoctoral research stay at IDEEAS in 2022-I.

Hidalgo, A., Gabaly, S., Morales-Alonso, Urueña, A. (2020). The digital divide in light of sustainable development: An approach through advanced machine learning techniques. *Technological Forecasting and Social Change*, 150.

Instituto Nacional de Estadística y Geografía (INEGI) (2020). *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH): Nota técnica*. INEGI.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). *An introduction to statistical learning*. Springer.

World Economic Forum (WEF) (2022). *World Economic Forum Annual Meeting Report*. World Economic Forum.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16466

Can websites reveal a firm's innovativeness? Empirical evidence on Italian manufacturing SMEs

Carlo Bottai[1], Lisa Crosato[2], Josep Domenech[3], Marco Guerzoni[1], Caterina Liberati[1]
[1]Department of Economics Management and Statistics, University of Milano-Bicocca, Italy, [2]Department of Economics, Ca' Foscari University of Venice, Italy, [3]Department of Economics, Universitat Politècnica de València, Spain.

Abstract

Research in innovation usually builds on conventional data such as balance sheets, surveys, patents, or product catalogs. This paper intends to explore unconventional data, specifically web-scraped data, as an information source for innovation studies, proposing a careful procedure to establish the veracity of the linkage between web-based data and firm-level information retrieved from conventional sources. The study regards a sample of Italian manufacturing small and medium enterprises active in 2016, comprehending both innovative and non-innovative firms. It is based on HTML tags, whilst most of the previous literature worked on the web-pages text and related semantics. Our paper provides evidence that the way HTML language is applied to build a corporate website unveils the capabilities of the owner firm, helping to distinguish innovative from non-innovative SMEs.

Keywords: *innovation, SMEs, unconventional data, HTML code, web-scraping*

1. Introduction

A firm's innovativeness surely ingrains its practices and the shared knowledge of its employees, but is not always observable. This complicates the assessment of the presence and intensity of innovative activity, although the innovation economics literature has made significant progress in measuring this phenomenon using balance sheets, surveys, patents, and product catalogs. None of these conventional sources, however, can completely capture such latent features, in particular as far as small and medium-sized enterprises (SMEs) are concerned (OECD [1963; 1992]). Furthermore, the derived innovation policy indicators are not always updated enough to describe the current situation. Balance sheets, for instance, are available at the end of each year and released by data providers with further delay.

This study suggests that SMEs' corporate websites, as outputs of a firm's activity, can represent an additional data source to build indicators of firms' innovative character. Firms typically shape their website as virtual showcases to sell products and share information related to their business. This makes the content of corporate websites highly connected to the economic activity of the firms [Domènech et al., 2012]. Moreover, websites are publicly accessible and regularly updated, so to appear as good candidates to solve some of the limitations of the currently available, conventional, sources. Accordingly, part of literature started scraping websites for research purposes [e.g. Blázquez et al., 2018; Crosato et al., 2021], and to use corporate websites to analyze firms' innovative activity [Libaers et al., 2016; Gök et al., 2015; Héroux-Vaillancourt et al., 2020; Daas and van der Doef, 2020; Kinne and Axenbeck, 2020; Axenbeck and Breithaupt, 2021; Kinne and Lenz, 2021, Ashouri et al., 2022].

Our paper adds to this literature but shifts the focus from the semantic analysis of web-pages text to the HTML code structure of webpages. The HTML code employed to build a corporate website stems from a blending of the company's needs and skills with those of the programmers [Brinck et al., 2001], so that it is sensible to suppose it unveils latent features such as high skills and creativity linked to the innovativeness of a firm. Innovative SMEs are supposed to be oriented towards new products development and commercialization, so we may expect that they want their websites well indexed by search engines and social networks. They employ high-skill workers, so they are keener to adopt new technologies, which should emerge from the HTML structure. Moreover, equipping a website with e-commerce, customer engagements, and user monitoring is easier through particular HTML programming styles. Finally, from a researcher's point of view, the analysis of a much more structured language like HTML is easier and computationally less expensive when compared to the analysis of natural languages applied in previous works.

2. Data Description

Our dataset merges conventional and unconventional data sources, where by *conventional* we refer to data resulting from a traditional design, i.e. originally collected by reference institutions for administrative purposes, but available in the standard matrix format and ready to be used for research purposes. Conventional data sources (Orbis and Aida databases by Bureau van Dijk) were the starting point to build the sample and were essential to divide the sample in innovative and non-innovative firms. Our unconventional source of data is the Wayback Machine of the Internet Archive (https://web.archive.org/). Table 1 summarizes the type of information retrieved from the different sources.

Table 1: Framework for dataset building

Data source and type	**Sample Units**	**Collected variables**
Orbis-BvD (conventional)	Italian Manufacturing SMEs, Active in 2016 with reported website N= 77,993	Sample selection variables - 'status' (active, bankrupt, in liquidation, etc.) - number of employees - total assets - turnover - website URL Company's details: - tax identification number (codice fiscale) - business name - business address (street name, number, and postcode) - telephone number Additional Stratification variables: - industrial sector (NACE) - geographical location (NUTS 2)
Aida-BvD (conventional)	Italian Manufacturing SMEs retrieved from Orbis	Label of innovative SME, as defined by the *Italian Startup Act*
Wayback Machine (unconventional)	Italian Manufacturing SMEs retrieved from Orbis with: - Website present in Wayback machine - Website ownership checked by our matching algorithm N= 43,335	HTML tags used to structure the website's front-page

Italian manufacturing SMEs, active in 2016, were retrieved from Orbis, using the standard definition of Eurostat based on the three firm size variables reported in Table 1. Firms with recorded websites were 77,993 on a total of 116,389. Since Orbis does not classify firms in terms of their innovativeness, we have resorted to its companion dataset Aida, which focuses on Italian firms. Here we have exploited the list of 'innovative' SMEs collected by the Italian Chamber of Commerce's Business Register in compliance with the Italian Startup Act (221/2012 law). Note that the Italian Startup Act has a few advantages with respect to other indicators of firms' innovativeness [Guerzoni et al., 2021]. It concentrates on SMEs who must focus on novel products and it does not base the classification on a single innovation measure, so included firms possess at least one among the usual innovation proxies.

The collection of the web-based information required a previous rigorous screening process to assess the attribution of the website reported in Orbis for each firm, to be reasonably sure that we are observing the true website of the company of interest. This is a fundamental aspect of our sampling design that, to the best of our knowledge, was not taken care for but in a few works (such as Barcaroli et al. [2016]). To this end, we have accessed the 2016 archived version of each firm's website URL, as reported in Orbis, on the Wayback Machine and, in each of the reported websites, we have searched for the company's details collected from Orbis, both in the front-page of the website and in any of the web-pages reachable from a hyperlink contained in the front-page.

At the end of this process we were left with 43,335 SMEs, after removing firms whose 2016 was not archived in the Wayback Machine (about 14%) and firms whose website was not confirmed as their own by our algorithm (about 30%, including websites in which the details of the firms were not available).

Our unconventional, web-based, information was finally scraped from the SMEs verified corporate website, with a procedure similar to the one described in Blázquez et al. [2018] and Crosato et al. [2021]: we have accessed the websites front-page and collected any HTML tag used to structure the page. We have kept only the HTML tag used more than three times in the whole document corpus, thus obtaining a final set composed of 711 HTML tags and six aggregate web-based statistics.

3. Can SMEs websites unveil the innovative character of its owner?

In our sample, SMEs labeled as "innovative" in the Aida dataset were only 178. In order to assure a fair comparison between the two groups SMEs, we structured 100 samples of 680 non-innovative firms stratifying by firm size, region and industry to match the smallest sample of innovative firms.

To spot the difference in the HTML structure of innovative and non-innovative corporate websites, each of the collected descriptive statistics and HTML tags were then compared on each of the one hundred non-innovative samples against the group of innovative SMEs.

Results suggest that corporate websites of innovative SMEs appear to be bigger either when measured by the HTML code underlying their front-page, and by the embedded text as measured by the variables reported in Table 2. The variables *text_size* and *gztext_size* represent the amount of text used on the page, but the latter does not take into account repetitions; *html_size* measures the size of the HTML code of the web-page analyzed; *href_number* and *img_number* represent the number of hyperlinks and images present on the page, respectively. Finally, *linkhref_number* counts the number of external resources used by the page. The hypothesis of the two samples being drawn from the same distribution is strongly rejected by both the Kolmogorov-Smirnov (KS) and the Mann-Whitney tests: almost all the considered features show median p-values (Table 2) under the 5% significance level. We can clearly see that the size distributions of the HTML and the text shift to larger values for innovative firms with respect to controls (Figure 1).

Table 2: P-value of Kolmogorov–Smirnov (KS) and Mann–Whitney (MW) tests comparing the distribution of web-based variables (innovative SMEs VS one hundred samples of control firms, median pvalue)

Variable	KS-test	MW-test
text_size	0.003	0.001
gztext_size	0.002	0.001
html_size	0.001	0.000
href_number	0.001	0.000
img_number	0.083	0.012
linkhref_number	0.000	0.000

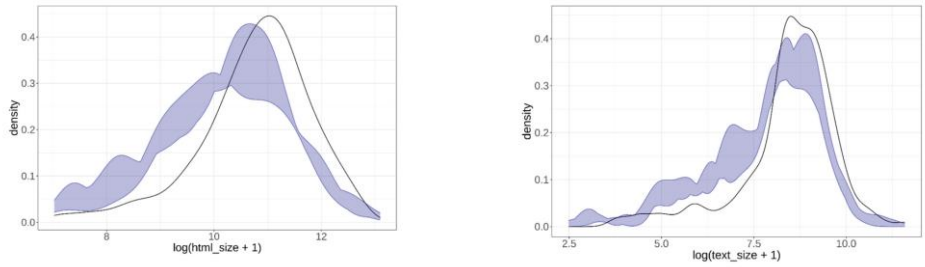


Figure 1: Density distributions of two descriptive statistics about the web-pages for innovative (solid line) and non-innovative (lilac fan on one hundred samples) SMEs.

As about the HTML tags used in the source code of the front-pages, several of them seem to discriminate between innovative SMEs and control firms. In this case, we have added the $\chi 2$ test for independence between belonging to the innovative group and presence of the feature, since for a few tags the KS test and the MW test disagreed. In Table 3 we have grouped the tags according to whether they discriminate in conformity with two out of three tests (moderate discriminating power) or with all of the three (highly discriminating power). We reject the null of similarity or independence if the median p-value of the test repeated over the 100 samples is smaller than 5%.

Table 3: HTML tags grouped according to moderate (2 out of three test) or high (3 out of 3 tests) discriminating power.

Number of test rejecting the hypothesis of similarity/independence	**TAGS**
Two out of three	<a>, , <embed>, <link>, <meta>, <nav>, <object>, <p>, <param>, <script>, <section>, <style>
Three out of three	<footer>, <header>, <h>, <i>, , , <table>, <td>, <tr>,

Among the first group of tags, a few of them are of the kind essential for websites building: , <a>, <p>, independently of the degree of innovativeness of the firm. On the other way round, highly discriminating tags include <footer> or <header>.

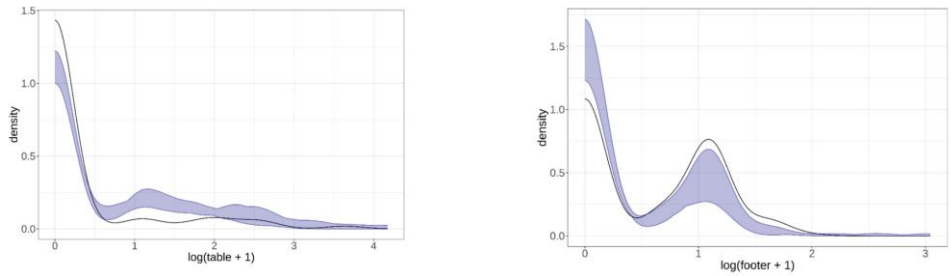


Figure 2: Density distributions of the number of times the indicated tag is used in a web-page for innovative (solid line) and non-innovative (lilac fan on one hundred samples) SMEs.

The distribution of occurrences of the former tag is represented in the right panel of figure 2 and shows that <footer> is more frequent in the front-pages of the websites of innovative SMEs. Tags like <table> (Figure 2, left panel) or <embed> are instead less used by innovative SMEs. Tables are nowadays deprecated, in favor of alternatives tailored for mobile devices.

The tag <embed>, mostly used to include Adobe Flash content in web-pages, is less present in the innovative SMEs web-pages since Adobe Flash was gradually set aside since the early 2010s. These examples confirm that corporate websites of the innovative SMEs rely more on the modern HTML (HTML5) with respect to those of comparable non-innovative firms.

4. Conclusions

In this paper, we have proposed and explored the use of unconventional data scraped from corporate websites as a complementary source of information for identifying innovative SMEs. Our results point out some of the characteristics shaping either group of firms: we found bigger websites and more updated HTLM language for the innovative SMEs group. These findings, although preliminary, confirm the underlying hypothesis that the HTML code of corporate websites and its characteristics represent observable proxies high skills and ingeniousness, characterizing the ability of an SME to embrace innovation. We thus provide the first contribution trying to translate the HTML code of corporate websites into data for identifying innovative firms and derive innovation policy indicators, relatively inexpensive to build and easy to be constantly updated. Ongoing research pursues an unsupervised learning approach to understand whether a natural grouping of the HTML tags emerges from the data.

Acknowledgments

We thank the Italian Ministry of University and Research (MUR) for sponsoring this work under the 'Departments of Excellence 2018-2022' funding schema, and the DEMS Data Science Lab of the University of Milano–Bicocca for computational resources. Josep Domenech acknowledges that this research was partially funded by MCIN/AEI/10.13039/501100011033 under grant PID2019-107765RB-I00.

References

Ashouri, S., A. Suominen, A. Hajikhani, L. Pukelis, T. Schubert, S. Türkeli, C. Van Beers, and S. Cunningham (2022). "Indicators on firm level innovation activities from web scraped data". Data in Brief, 42:108246.

Axenbeck, J. and P. Breithaupt (2021). "Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity?". PLOS ONE, 16(4):1–23.

Barcaroli, G., M. Scannapieco, and S. Donato (2016). "On the use of Internet as a data source for official statistics: A strategy for identifying enterprises on the Web". Rivista Italiana di Economia Demografia e Statistica, 70(4):25–41.

Blázquez, D., J. Domènech, and A. Debón (2018). "Do corporate websites' changes reflect firms' survival?" Online Information Review, 42(6):956–970.

Brinck, T., D. Gergle, and S. D. Wood (2001). Usability for the Web: Designing Web Sites that Work. Elsevier.

Crosato, L., J. Domènech, and C. Liberati (2021). "Predicting SME's default: Are their websites informative?" Economics Letters, 204:109888.

Daas, P. J. H. and S. van der Doef (2020). "Detecting innovative companies via their website". Statistical Journal of the IAOS, 36(4):1239–1251.

Domènech, J., B. de la Ossa, A. Pont, J. A. Gil, M. Martinez, and A. Rubio (2012). "An intelligent system for retrieving economic information from corporate websites". In IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 573–578.

Gök, A., A. Waterworth, and P. Shapira (2015). "Use of web mining in studying innovation". Scientometrics, 102(1):653–671.

Guerzoni, M., C. R. Nava, and M. Nuccio (2021). "Start-ups survival through a crisis. Combining machine learning with econometrics to measure innovation". Economics of Innovation and New Technology, 30(5):468–493.

Héroux-Vaillancourt, M., C. Beaudry, and C. Rietsch (2020). "Using web content analysis to create innovation indicators—What do we really measure?". Quantitative Science Studies, 1(4):1601–1637.

Kinne, J. and J. Axenbeck (2020). "Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study". Scientometrics, 125(3):2011–2041.

Kinne, J. and D. Lenz (2021). "Predicting innovative firms using web mining and deep learning". PLOS ONE, 16(4):1–18.

Libaers, D., D. Hicks, and A. L. Porter (2016). "A taxonomy of small firm technology commercialization". Industrial and Corporate Change, 25(3):371–405.

OECD (1963). Frascati Manual: The Proposed Standard Practice for Surveys of Research and Experimental Development. OECD Publishing.

OECD (1992). Oslo Manual: OECD Proposed Guidelines for Collecting and Interpreting Technological Innovation Data. OECD Publishing.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16495

Digital transformation strategies for the sustainable growth of startups in Australia

Majdah AL Nefaie[1], Siva Muthaly[2], Shahadat Khan[3]
[1]General Manager of Marketing and insights, Saudi Arabia, [2]School of Business, western Sydney University, Australia, [3]School of Business and IT technology, University of Royal Melbourne technology, Australia

Abstract

The COVID-19 outbreak provided a glimpse of a future world in which digital interactions are critical, compelling organisations and individuals to increase their adoption of technology. Therefore, to gain a competitive edge, organisations need accurate, real-time responses from extensive data analysis to further develop their products and services or create entirely new business models. The rising dependence on technical services has created a gap between organisational offerings and objectives, especially for startups. Statistics reveal the high failure rate of startups in top-ranking countries such as the United States, Germany and Australia. A major reason for startup failure is a lack of business knowledge in the technical team, affecting the adaptability of the business with respect to technology. Australia is ranked eighth in the world in terms of its startup ecosystem because of its internal market dynamics and physical, commercial and legal infrastructure. Nevertheless, startup failure in Australia is significant. Therefore, this paper, based on desk research and an analysis of secondary data, presents digital transformation strategies aimed at reducing risks, creating sustainable business growth and providing a real-time view of business efficiency, resulting in reduced costs and improved performance.

Keywords: *Startups; digital transformation; bid data; Australia; economy; sustainable business growth.*

1. Introduction

This research paper discusses the effect of digital transformational strategies on the sustainable growth of startups and the national economy. The paper provides an overview of the global startup market before focusing specifically on the Australian market. Given its internal market dynamics and physical, commercial and legal infrastructure, Australia ranks eighth in the world for startups (Statista, 2022).

The COVID-19 pandemic has compelled businesses across a wide range of sectors to provide online services, and consumers have become dependent on these services for their essential and non-essential needs. Under these new societal norms, business demand for cloud- and internet-based technologies is increasing (Hai, Van, & Thi Tuyet, 2021). The increasing adoption of digital solutions for existing services or operations across industries has boosted productivity (Osmundsen, Iden, & Bygstad, 2018). In 2022, the global digital transformation market was valued at US$594.5 billion and is predicted to reach US$1,548.9 billion by 2027, showing a compound annual growth rate (CAGR) of 21.1% over the forecast period (Statista, 2022). To gain a competitive edge, organisations need digital solutions to gain real-time responses from extensive data analysis to create new or enhance existing products and services or completely recreate their business models (Hai et al., 2021).

The increasing dependence on technical services has created a gap between organisational offerings and business objectives, especially for startups. A lack of business knowledge in the technical team can affect the adaptability of the organisation with respect to technology (Savey, Daradkeh, & Gouvela, 2020). Studies show that, globally, 90% of startups fail, with 20% failing in the first year, 30% failing within 2 years, and 50% failing within 5 years (Statista, 2022). The predominant reasons for startup failure include technological issues, lack of skills and poor financial capability (Savey et al., 2020).

The gap between business and technical knowledge means that potential technological solutions such as big data integration systems, open data analysis, public data mining, artificial intelligence (AI) and cybersecurity strategies offer no significant benefits for businesses (Hai et al., 2021). More research on the organisational need for such technological solutions is needed before technical services can be effectively provided (Sekongo, 2019).

Cybercrime involving the theft of sensitive client data and financial information has become a major risk for businesses (Statista, 2022). This has led to the increased demand for secure digital transformation solutions across various industries. Therefore, this paper

presents high-end digital solutions in the form of digital transformation strategies for startups to help improve efficiency and productivity, reduce operational costs and increase net profits through digitisation and automation. These strategies are aimed at reducing risks and creating a safer, healthier and more sustainable business with a real-time view of efficiency to reduce costs and improve performance.

This study is based on a desk research methodology in which data were gathered from recent research articles and statistical resources. These data reveal the impact of digital transformation on the global and Australian economies.

The remainder of the paper is structured as follows: Section 2 presents the statistics related to the global startup market, the success and failure rate of startups by country and the gaps in the Australian market; Section 3 discusses the e-services market in Australia; Section 4 presents digital transformation strategies for startups to ensure their sustainability; and Section 5 provides conclusions and recommendations.

2. Global Startup Market Statistics

According to data in figure 1, the United States (US) was the top-ranked country for startups in 2022 , with a score of nearly four times (195.37) that of the second top-ranked ranked country, the United Kingdom (52.56). These ranking are based on the country positive ecosystem and number of established startups. Australia ranked eighth, placing it strongly in the top 10 leading countries for startups worldwide.

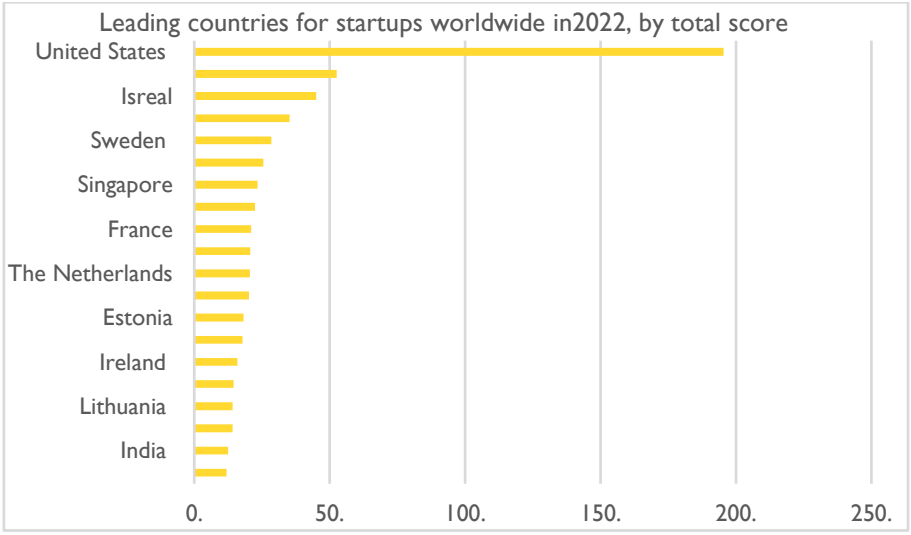


Figure 1: Leading countries for startups worldwide 2022, (Statista 2022).

Australia's strong ranking arises from its internal market dynamics and physical, commercial and legal infrastructure (Statista, 2022), which have strongly influenced its startup ecosystem and resulted in it becoming one of the first countries to establish startups.

2.1. Startup Success and Failure Rate by Country

Statistics show that startup failure rates are higher than success rates (see Table 1). South Africa has the highest startup failure rate at 86% and the lowest success rate at 14%, while Switzerland has the lowest failure rate at 65% and highest success rate at 35%. The US, Canada and France have similar startup failure and success rates of 80% and 20%, respectively. Australia, Germany and Estonia all have a 75% failure rate and a 25% success rate. These statistics show that startup failure rate is high in all top-ranked countries.

Table 1: Startup Success and Failure Rate by Country

Country	Startup Failure Rate	Startup Success Rate
United States	80%	20%
Canada	80%	20%
France	80%	20%
Germany	75%	25%
Switzerland	65%	35%
Estonia	75%	25%
South Africa	86%	14%
Australia	75%	25%

Source: CB Insights. (2023).

The most common reason for startup failure is a lack of product demand arising from a limited understanding of customer needs (Savey et al., 2020). Moreover, startups in the technological domain have high failure rates; for example, 80% of e-commerce, 75% of fintech, 80% of health technology and 60% of educational technology startups fail (CB Insights, 2023).

2.2. Gaps in the Australian Market

Challenges in the Australian startup market contribute to startup failure within the first 3 years. The primary reasons for startup failure include financing issues, finding the right talent and navigating the regulatory and legal environment (Expert-Market, 2023). Startups need to secure resources and implement cost-effective operations to minimise their financial obligations. Recruiting skilled people for startup teams can be costly; therefore,

outsourcing may be an effective way to collaborate with professionals at a lower cost. Navigating the regulatory and legal environment is challenging for startups in Australia, especially with respect to digital services and data security. Startup owners must understand the relevant laws and regulations pertaining to their online services and the potential implications of non-compliance. Therefore, startups should consult with technical business advisors to ensure their compliance with relevant legislation.

3. The E-Services Market in Australia

Revenue from the e-services market is predicted to reach US$5.40 billion in 2023 and show a CAGR of 8.71% (2023–2027), resulting in a projected market value of US$7.54 billion by 2027 (Statista, 2022). Figure 2 shows that revenue from the information technology (IT) services market is projected to reach US$34.11 billion in 2023. The largest segment in the market is IT outsourcing, with a projected market volume of US$12.07 billion in 2023. This revenue is expected to show a CAGR (2023–2027) of 7.06%, resulting in a market volume of US$44.81 billion by 2027. The average spend per employee in the IT services market is projected to reach US$2,470 in 2023.

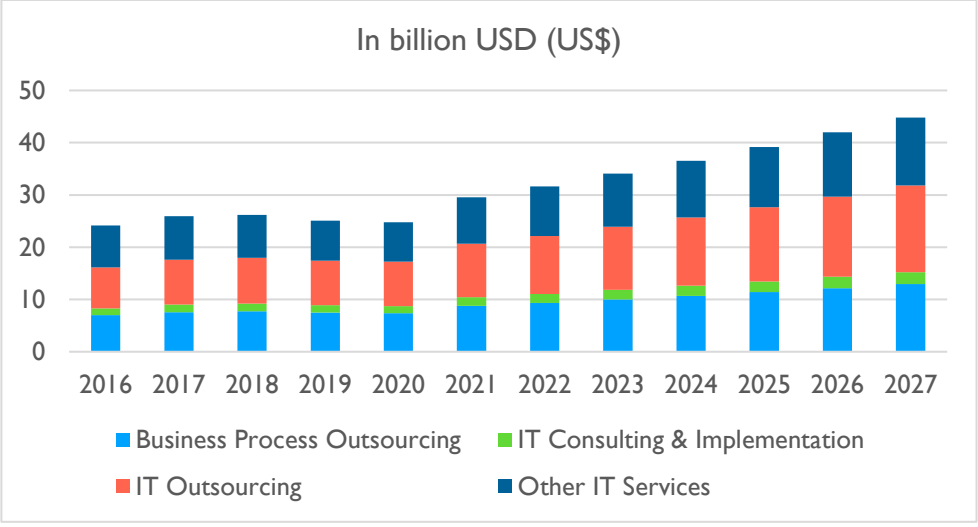


Figure 2: 3. The E-Services Market in Australia 2022, (Statista, 2022).

Online education and training are the largest segments in the market, with an expected revenue growth of 14.4% by 2024. The education and training sector provides educational services through preschools, primary and secondary schools, technical colleges, training centres and universities (Statista, 2022).

The growth of e-services and the dependence on big data with respect to consumer transactions and social media mean that startups face major challenges, which can lead to failure. Therefore, to grow their businesses sustainably, startups need the support of technology professionals (Hai et al., 2021). The following section presents digital transformational strategies for technology professionals to assist startups with sustainable business growth.

4. Digital Transformational Strategies for Sustainable Startup Growth

Digital transformation can enhance performance and reduce errors, saving time and effort. In addition, it increases the flexibility, convenience, collaboration and performance of organisational activities (Nguyen, 2020). Therefore, digital transformation has become essential for both small and large businesses to enhance their business processes and customer experiences to meet changing business requirements.

Startups are playing an increasingly important role in the Australian economy. With the growth in technology, cyberthreats and cybercrime have taken on new shapes in the form of next-generation ransomware, web attacks and others (Expert-Market, 2023). This paper presents digital transformational strategies for technology professionals who offer services to startups that require technological solutions for services such as online shopping, open and public data mining, shipping logistics, online systems, video conferencing, AI, virtual reality and interactive systems.

The business strategies presented in this paper are based on the need for high-end digital solutions, consultation and training. The primary objective is to provide startups with limited resources in the rapidly growing e-services and online business sectors with secure, cost-effective digital solutions to ensure sustainable growth and meet client needs. Startups seek digital solutions to create better products more quickly and at a lower cost (Savey et al., 2020). These solutions can help startups be more sustainable, data driven and compliant (El Hilali & El Manouar, 2019). Figure 3 illustrates various IT solutions that may be offered to startups to make the crucial first steps and confidently expand their businesses. These solutions include integrated platforms with advanced and cost-effective technological features to boost the customer base, training and consultation in digital services, cybersecurity and compliance with legislation, strong data encryption practices, the efficient deployment of resources to improve decision-making and reduce operational costs and the offering of digital service packages at prices that meet the startup’s financial capabilities.

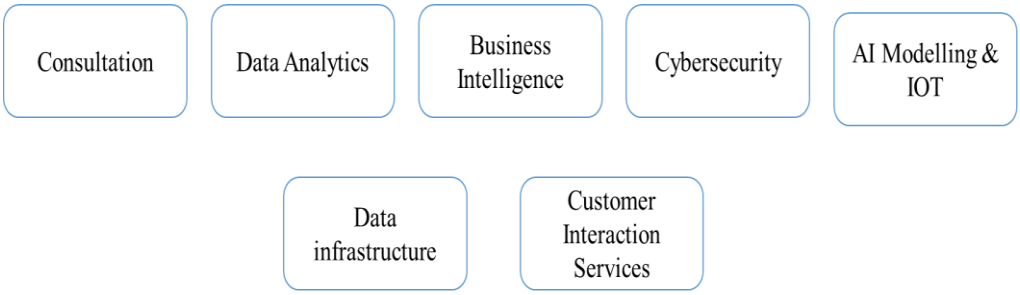


Figure 3: Services that may assist startups.

5. Conclusion

This paper has presented digital transformation strategies for technology professionals to assist startups with sustainable business growth and meeting client needs in the Australian market. Technology professionals can offer products and services such as training and consultation, data analytics, business intelligence, cybersecurity checks, AI modelling and systems, the Internet of Things, data infrastructure and customer interaction services to help startups find technological solutions for services such as online shopping, open and public data mining, social media analytics, video conferencing, AI, virtual reality and interactive systems (Hai et al., 2021). In addition, the paper presented strategies for technology professionals to assist startups to make the best possible use of IT, such as training and consultation in utilising digital services, applying secure digital solutions, managing cybersecurity systems and implementing strong data encryption practices.

The statistics presented in this paper reveal that startup failure rates are higher than success rates in leading startup countries. For example, the challenges faced by startups in the Australian market often lead to failure within the first 3 years. The main reasons for startup failure include financial issues, the lack of skilled people and an inability to navigate the regulatory and legal environment. These challenges have increased with the growth of e-services and the dependence on big data pertaining to consumer transactions and social media, contributing to the failure of startups across various sectors. Therefore, it is recommended that startups seek the help of technology professionals to grow their businesses sustainably. Studies have shown that the adoption of various advanced technologies such as AI, machine learning, the Internet of Things, blockchain technology and 5G have a positive effect on the market, resulting in the exponential growth of businesses across the globe (Hai et al., 2021). In addition, investments in the IT sector and the development of smart applications can lead to a significant increase in the adoption of big data and other related technologies. Therefore, to enhance the adoption of technology and utilisation of big data, it is critical to deploy simple, cost-effective, advanced digital

transformational strategies to promote the sustainable growth of startups and reduce the risk of failure.

References

CB Insights. (2023). Retrieved from https://www.cbinsights.com/

El Hilali, W., & El Manouar, A. (2019). Towards a sustainable world through a SMART digital transformation. In B. Abouelmajd & M. Ben Ahmed (Eds.), *Proceedings of the 2nd International Conference on Networking, Information Systems & Security* (pp. 1–8). New York, NY: The Association for Computing Machinery. https://doi.org/10.1145/3320326.3320364

Expert-Market. (2023). Examining the top challenges faced by start-ups in Australia. Retrieved from https://www.expert-market.com/examining-the-top-challenges-faced-by-start-ups-in-australia/

Hai, T. N., Van, Q. N., & Thi Tuyet, M. N. (2021). Digital transformation: Opportunities and challenges for leaders in the emerging countries in response to Covid-19 pandemic. *Emerging Science Journal, 5*, 21–36. https://doi.org/10.28991/esj-2021-sper-03

Nguyen, T. A. (2020). Thúc đẩy chuyển đổi số tại Việt Nam [Promoting digital transformation in Vietnam]. *Vietnam Journal of Science and Technology*. Retrieved from https://vjst.vn/vn/tin-tuc/3302/thuc-day-chuyen-doi-so-tai-viet-nam.aspx

Osmundsen, K., Iden, J., & Bygstad, B. (2018). Digital transformation: Drivers, success factors, and implications. Paper presented at the 12th Mediterranean Conference on Information Systems (MCIS 2018), Corfu, Greece. Retrieved from https://aisel.aisnet.org/mcis2018/37

Ruby, D. (2023). *107+ startup statistics for 2023 (global facts and figures)*. Retrieved from https://www.demandsage.com/startup-statistics/

Savey, L., Daradkeh, Y. I., & Gouvela, L. B. (2020). The success of startups through digital transformation. *International Journal of Open Information Technologies, 8*(5), 53–56.

Sekongo, M. (2019). Digital: Transformation digitale, échec ou succès [Digital transformation, failure or success]. *Strat'Marques*. Retrieved from https://stratmarques.com/digital-transformation-digitale-echec-ou-succes/

Statista. (2022a). Cybersecurity—Australia. Retrieved from https://www.statista.com/outlook/tmo/cybersecurity/australia

Statista. (2022b). IT services—Australia. Retrieved from https://www.statista.com/outlook/tmo/it-services/australia

Statista. (2022c). Leading countries for startups worldwide in 2022, by total score. Retrieved from https://www.statista.com/statistics/1275240/leading-countries-startups-worldwide/https://www.statista.com/outlook/tmo/it-services/australia

SDG 9 - Industry, Innovation and Infrastructure: Impact on the digital sphere discussion

Enara Zarrabeitia-Bilbao[1], Rosa María Rio-Belver[1], Maite Jaca-Madariaga[1], Izaskun Álvarez-Meaza[1]
[1]Industrial Organization and Management Engineering Department, University of the Basque Country, Spain.

Abstract
The research carried out analyzes more than 5 million tweets on Sustainable Development Goals (SDGs), in general, and more than 17,000 tweets on SDG 9 (Industry, Innovation and Infrastructure) , in particular, in the three-year period 2020-2022. After using Social Network Analysis and Semantic Analysis techniques, the results obtained show that SDG 9 has generated less interest in the social network Twitter than the other SDGs in the last three years, and the number of tweets about it has been decreasing. Moreover, the private sector does not play a key role in any of the main communities generated. Nevertheless, it is considered necessary to improve the communication strategy of SDG 9 and make it mainstream.

Keywords: *Sustainable Development Goals; SDG 9; Twitter, Social Network Analysis; Semantic Analysis.*

Deepening big data sustainable value creation: Exploring the IPMA and NCA perspectives

Randy L. Riggs[1], Carmen M. Felipe[1], José L. Roldán[1], Juan C. Real[2]
[1]Department of Business Administration and Marketing, Universidad de Sevilla, Spain,
[2]Department of Business Management and Marketing, Pablo de Olavide University, Spain

Abstract

The impact of big data analytics capabilities (BDAC) on firms' sustainable performance (SP) is exerted through a set of underlying mechanisms that operate as a 'black box.' Our previous research demonstrated that a serial mediation of supply chain management capabilities (SCMC) and circular economy practices (CEP) is required to improve SP from BDAC. However, further insights regarding the role of BDAC in the processes of SP creation can be provided by deploying complementary analytics techniques, namely the importance-performance map analysis (IPMA) and the necessary condition analysis (NCA). This paper runs these techniques on a sample of 210 Spanish companies with the potential for circularity and environmental impact. The results show that BDAC are essential for achieving SP. However, companies still have enough room for improvement to take further advantage of these capabilities. Additionally, BDAC are a necessary (must-have factor) and sufficient (should-have factor) condition for the rest of the variables in the model. Furthermore, high levels of BDAC are required to achieve excellence in SP.

Keywords: *Big data analytics capabilities; Circular economy practices; Supply chain management capabilities; Sustainable performance; Importance-performance map analysis (IMPA); Necessary condition analysis (NCA).*

Data frequency and forecast performance for stock markets: A deep learning approach for DAX index

Diana A. Mendes[1], Nuno B. Ferreira[1], Vivaldo M. Mendes[2]
[1]Iscte-University Institute of Lisbon and BRU-IUL, Department of Quantitative Methods for Management and Economics, Portugal [2]Iscte-University Institute of Lisbon and BRU-IUL, Department of Economics, Portugal

Abstract

Due to non-stationary, high volatility, and complex nonlinear patterns of stock market fluctuation, it is demanding to predict the stock price accurately. Nowadays, hybrid and ensemble models based on machine learning and economics replicate several patterns learned from the time series.

This paper analyses the SARIMAX models in a classical approach and using AutoML algorithms from the Darts library. Second, a deep learning procedure predicts the DAX index stock prices. In particular, LSTM (Long Short-Term Memory) and BiLSTM recurrent neural networks (with and without stacking), with optimised hyperparameters architecture by KerasTuner, in the context of different time-frequency data (with and without mixed frequencies) are implemented.

Nowadays great interest in multi-step-ahead stock price index forecasting by using different time frequencies (daily, one-minute, five-minute, and ten-minute granularity), focusing on raising intraday stock market prices.

The results show that the BiLSTM model forecast outperforms the benchmark models –the random walk and SARIMAX - and slightly improves LSTM. More specifically, the average reduction error rate by BiLSTM is 14-17 per cent compared to SARIMAX. According to the scientific literature, we also obtained that high-frequency data improve the forecast accuracy by 3-4% compared with daily data since we have some insights about volatility driving forces.

Keywords: *Time Series Prediction, SARIMAX model, LSTM and BiLSTM model, German stock market.*

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16472

Analysis of the effectiveness of measures to reduce the severity of traffic accidents in the city of Barcelona in the period 2013-2019

Lluís Bermúdez[1], Isabel Morillo[1]
[1]Department of Economic, Financial and Actuarial Mathematics, Universitat de Barcelona, Spain

Abstract

We study the severity of traffic accidents in the city of Barcelona during the period from 2010 to 2019. We intend to measure the performance of Local Road Safety Plan in Barcelona 2013-2018 actions in reducing fatal/ seriously injured victims, throughout the whole period examined. We also analyse the effect of the risk factors on accident severity to detect which are significant and therefore on which the following measures can be focused to reduce severity. We draw on data available on the Open Data Barcelona platform. Logistic regression model is applied. The results show that the 2013-2016 period presents a lower risk of fatal/serious injuries with a reduction of the severity odds ratio in 10%. This lower risk is even greater in the 2017-2019 period, with a reduction of 18%. In general terms would confirm that the measures have had an effect. Furthermore, it can be observed that there has been a reduction in severe accidents on working days, as well as on the day shift and on normal streets. On the other hand, the incidence of severe accidents has remained the same when two-wheeled or heavy vehicles are involved and when there is speeding, a run-over or a shock.

Keywords: *Road traffic accidents; risk factors; road safety.*

1. Introduction

The analysis of traffic accidents and the adoption of measures to reduce them is on the agenda of the European cities. The Local Road Safety Plan (LRSP) in Barcelona 2013-2018 **(Ajuntament de Barcelona, 2013)** is an initiative which in the technical field, focuses on preventive and corrective actions as well as education for safe mobility with a specific objective: reduce by 30% the number of deaths in traffic accidents and 20% of serious injuries in traffic accidents in 2018 compared to 2012 levels.

In this study we intend to measure the performance of LRSP measures in reducing fatal and seriously injured victims throughout the whole period examined. The severity of road accident casualties in urban areas and the analysis of measures taken for their reduction have been investigated in different studies (see e.g. Manner & Wünsch-Ziegler (2013); Wang *et al* (2019)). For our purposes, we use a data set with detailed information on accidents with victims that occurred in Barcelona, for the years 2010 until 2019. We evaluate the effect of the year in which the accident occurred into crash-injury severity, controlling for a number of different risk factors including type and cause of accident, number and type of vehicles involved, type of day and type of road where accident occurred. In this way, we can detect which are significant and therefore on which the following measures can be focused to reduce severity. For this purpose, we have divided the years in which accidents occurred into 3 periods: the period 2010-2012, in which the characteristics of the accident rate is taken as a reference for the design of the LRSP; the period 2013-2016, when the measures started to be implemented; and the last period 2017-2019, which coincides with the completion of the plan. To examine the contribution of the risk factors considered to the severity of the accidents, we have categorised traffic accident injury severity into fatal/serious and non-serious (minor or slight) and we have fitted a binary logistic regression model.

2. Data and methods

Traffic accident data were obtained from the database of traffic accidents managed by the local police in the city of Barcelona. The final database obtained include a total of 87,823 traffic accidents with victims during the studied period. The victims are classified into three categories: fatal, seriously injured (hospitalized for more than 24 hours) and non-seriously injured (treated at the scene of the accident, in hospital emergency services or hospitalized for less than 24 hours).

In this study, the modelling effort has been restricted to using only the predicting factors that reflect the crash/accident characteristics including accident time (*Period*, *Day* (of the week) and *Dayparts* (time of the day)); crash site (*Via* (street, avenue or fast lane)); related causes (*Pedestrian*, *Bloodalcohol*, *Speed* and *Roadcondition*); type of accident (*Runover*, *Twowheelscrash*, *Collision* and *Shock)*; the vehicle characteristics (*Vehicles* (number) and

type of vehicle involved (*Bicycle*, *Twowheels*, *Heavy* and *Light*). The study factors period, day, day parts and via were classified into several categories, the first category is the reference. In the rest of the factors, it has been necessary to convert them into dummy dichotomous variables. This potentials predicting factors have been extensively used in several studies (Vorkov-Jovic *et al.* (2006); Yau *et al*. (2006); Wang *et al*. (2019)).

To determine the associations between the probability of severity outcomes (e.g. fatal/serious injuries versus minor injuries) and all contributory factors, we fit a binary logistic regression ((Sze & Wong, 2007); Moudon *et al*. (2011)). The response variable is *Severity* with two levels (0: non-serious; 1: fatal and serious), the reference category is 'non-serious injury'. The logistic regression model, expressed in terms of the logit transformation of the ith individual's response probability, p_i (e.g. probability of fatal/serious), is a linear function of the vector of explanatory variables:

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_n x_n \qquad (1)$$

3. Results

Table 1 shows the results of binary logistic regression model estimated:

Table 1. Results of modelling the severity of the accidents.

	Logistic		
Variables	Coef.	SE.	P-val.
Intercept	-5.059	0.134	<0.01
Period_2013-16	-0.102	0.050	0.041
Period_2017-19	0.198	0.055	<0.01
Day_Working	-0.295	0.051	<0.01
Dayparts_Afternoon	0.215	0.047	<0.01
Dayparts_Even.-Night	0.397	0.070	<0.01
Via_Avenue	0.077	0.045	0.085
Via_Fast lane	0.246	0.087	<0.01
Vehicles_1	-0.947	0.117	<0.01
Bycicle_1	0.773	0.104	<0.01
Twowheels_1	1.249	0.076	<0.01
Heavy_1	1.359	0.093	<0.01
Light_1	0.421	0.079	<0.01
Pedestrian_1	0.393	0.075	<0.01
Bloodalcohol_1	0.157	0.115	0.172
Speed_1	1.314	0.132	<0.01
Roadcondition_1	-0.904	0.342	0.013
Runover_1	1.491	0.099	<0.01
Twowheelcrash_1	0.208	0.085	0.015
Collision_1	0.706	0.108	<0.01
Shock_1	1.477	0.092	<0.01
AIC			5958.4

Source: Open Data BCN. https://opendata-ajuntament.barcelona.cat/data/es/dataset.
Own elaboration.

The reported results are the estimated coefficients for the risk factors computed and the standard error and the p-value statistic for each variable. The Akaike information criterion (AIC) is also included as an indicator of the goodness of fit of the model.

Results show that several factors are found to be associated with the injury severity. In the case of the period in which the accident occurs, assumed as an indicator of the performance of the LRSP measures, period 2013-2016 presents a lower risk of fatal/serious injuries with a reduction of the severity odds ratio in 10%. This lower risk is even greater in the period 2017-2019, with a reduction of the severity odds ratio in 18%, which in general terms would confirm that the measures have had an effect.

For the rest of risk factors, the severity odds ratio is reduced by 25% during working days and is increased in 49\% when the accident occurs in the evening-night with respect to the morning shift. Setting normal streets as the reference category there is a higher risk of a fatal/serious injury in fast lanes, with an increase of 28% in the severity odds ratio. There is a lower risk of fatal/serious injury if there is more than one vehicle involved in the accident, the odds ratio is reduced in 61% with respect to an accident with only one vehicle involved. If two-wheeled or heavy vehicles are involved there is a higher risk of fatality/serious injury than if a light vehicle is involved. For example, when a two-wheeled is involved in the accident the severity odds ratio increase 3.5 times with respect the case in which is not. Speeding increases the severity odds ratio in 3.7 times. Excess alcohol is not such a significant factor and this is probably due to collinearity with excess of speed and night. Road condition does not increase the risk of deaths/serious injuries. Finally, the results show that run over and shocks (crashes against a static element) increase more the risk of accidents with fatalities and serious injuries. In both cases, the severity odds ratio increases around 4 times. These results are consistent with those of other studies ((Valent, et al., 2002); (Barua & Tay, 2010); (Manner & Wünsch-Ziegler, 2013); (Nasri et al. (2022)).

4. Conclusions

From the evaluation of the effect of the year in which the accident occurred into crash-injury severity, the results obtained in our study show that in general terms the severity (fatal/serious injury) of road accidents has decreased in the period 2013-2019. Considering the analysed risks factors, when two-wheeled or heavy vehicles are involved continue to increase the risk of a severe accident. This is also the case when the mediate cause is speeding and when the type of accident is a run-over or a collision with a static element. This casuistry has to be evaluated in greater detail within the accident rate, in order to seek out or influence in a more efficient way measures that help to achieve the objectives set.

References

Ajuntament de Barcelona. (2013). *Pla local de seguretat viària de Barcelona 2013-2018*. Barcelona. Retrieved from https://ajuntament.barcelona.cat/seguretatiprevencio/sites/default/files/PDF/PLASEGURETATVIARIABARCELONA2013-2108.pdf

Barua, U., & Tay, R. (2010). Severity of urban transit bus crashes in Bangladesh. *Journal of Advanced Transportation, 44*(1), 34-41.

Manner, H., & Wünsch-Ziegler, L. (2013). Analyzing the severity of accidents on the German Autobahn. *Accident Analysis and Prevention, 57*, 40-48.

Moudon, A., Lin, L., Jiao, J., Hurvitz, P., & Reeves, P. (2011). The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. *Accident Analysis and Prevention, 43*, 11-24.

Nasri, M., Aghabayk, K., Esmaili, A., & Shiwatoki, N. (2022). Using ordered and unordered logistic regressions to investigate risk factors associated with pedestrian crash injury severity in Victoria, Australia. *International Journal of Transportation Science and Technology, 81*, 78-90.

Sze, N., & Wong, S. (2007). Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. *Accident Analysis and Prevention, 39*(6), 1267-1278.

Valent, F., Schiava, F., Savonitto, C., Fallo, T., Brusaferro, S., & Barbone, F. (2002). Risk factors for fatal road traffic accidents in Udine, Italy. *Accident Analysis and Prevention, 34*(1), 71-84.

Vorko-Jovic, A., Kern, J., & Biloglav, Z. (2006). Risk factors in urban road traffic accidents. *Journal of Safety Research, 37*(6), 93-98.

Wang, D., Liu, Q., Ma, L., Zhang, Y., & Cong, H. (2019). Road traffic accident severity analysis: A census-based study in China. *Journal of Safety Research, 70*, 135-147.

Yau, K., Lo, H., & Fung, S. (2006). Multiple-vehicle traffic accidents in Hong Kong. *Accident Analysis and Prevention, 38*, 1157-1161.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16481

Redrawing electoral maps to curb gerrymandering: a case study of New York State in 2022

Shipeng Sun[1]
[1]Department of Geography and Environmental Science, Hunter College—The City University of New York, United States of America

Abstract

The delineation of electoral district boundaries is a fundamental component of democratic practice in the United States. However, gerrymandering—the manipulation of district boundaries to favor specific interest groups—undermines this process and often leads to contentious debates and legal battles. The primary objective of this study is to quantitatively evaluate four sets of New York State's 2022 congressional district maps for signs of gerrymandering. These maps were proposed by the Independent Redistricting Commission (IRC), the State Legislature, and the State Court, respectively. The quantitative metrics employed integrate factors such as population distribution, state boundaries, and spatial topology to assess district compactness and to identify gerrymandering. The results indicate that the Court-drawn congressional districts exhibit considerably lower levels of gerrymandering than the maps proposed by the IRC and the State Legislature, which exhibit little disparity. As the Supreme Court of the United States has ruled that addressing partisan gerrymandering falls within the jurisdiction of the state, the findings of this study suggest that appointing special map masters by the State Court and reducing or eliminating the influence of political parties in redistricting could generate fairer electoral maps that promote equitable representation of the state's populace.

Keywords: *redistricting; electoral maps; gerrymandering; New York State*

1. Introduction

Drawing electoral district boundaries is one fundamental component for the functioning of political systems in the United States (Crocker, 2012). Representatives of the House, for example, are elected every two years from the 435 congressional districts in the country. In November of even-numbered years, voters in each congressional district cast their ballots to elect their representative. The candidate who receives the most votes is elected to represent that district in Congress and voters outside the district have no direct impact on the election result. To reflect changes in population, every ten years, after the decennial census, states receives the numbers of the House representatives from apportionment and redraw their congressional district boundaries, as the constitution requires each district has roughly the same population.

Since district boundaries can be used to gather or dilute supporters of a particular political party or candidate, the undemocratic practice of gerrymandering, that is the manipulation of electoral district boundaries to favor particular group interests through "packing" and "cracking", proved to be an enduring challenge to eliminate, despite decades of efforts from political scientists, mathematicians, legal scholars, and engaged citizens (Abramowitz, Alexander, & Gunning, 2006; Ansolabehere & Snyder Jr, 2012). Even though racial gerrymandering against minorities has been ruled unconstitutional by the Supreme Court of the United States (SCOTUS) and therefore largely been prevented or corrected, partisan gerrymandering is still prevalent, partially because SCOTUS refused to judge the cases of partisan gerrymandering and suggested regulating district maps was the jurisdiction of the state. Although it can be reasonably argued that districting is political by nature and partisan gerrymandering also reflects, to a certain extent, the composition of the underlying constituents, extreme gerrymanders that lead to obviously weird-shaped boundaries suppress the representation of certain local communities and become a stain on the merit of democracy.

At the state level in the US, drawing electoral district maps has diverse practices, from governor-appointed committees to independent and third-party expert mapping groups, and to commissions approved by state legislatures. In the last two decades, the State of New York experienced different models of redistricting, particularly related to the 2022 mid-term elections (Table 1). The 2014 *New York Redistricting Commission Amendment* established the rule that a ten-member Independent Redistricting Commission (IRC) should be formed to redraw state legislative and congressional districts from 2021 onwards. Four legislative leaders each choose two commissioners, while the remaining two citizen-commissioners are selected by the eight members. The Commission shall submit proposed district maps to the Legislature, which can approve or reject the plans without modifications. The Legislature can only make amendments if the Commission's plans are rejected twice. Upon the release of the 2020 decennial census data, the IRC started working on redistricting maps in early 2021.

Divided by the party lines, however, the commission failed to reach consensus and had to submit Plan A and Plan B for the district maps. The State Legislature rejected both plans and the IRC could not submit a new plan within the required 15 days window. As a result, the legislature created its own maps, and the governor signed them into law. After that, the district maps of Congress and State Senate faced lawsuits and were struck down by the State Court of Appeals in April 2022. In the end, those two maps were drawn by the special master appointed by the Court.

Table 1. Timeline of Redistricting for 2022 Elections in New York State, USA

Time	Event
2014	New York State enacted a constitutional amendment to form the Independent Redistricting Commission (IRC) to draw district maps for congress, state senate, and state assembly.
Early 2021	Upon the release of new decennial census data, the IRC started to work on the new district maps.
Dec, 2021	The IRC could not reach consensus on the proposed district maps, with irreconcilable division between the Democratic and Republican party lines.
01/03/2022	The IRC submitted two separate plans for Congress, State Senate, and State Assembly, Plan A and Plan B favored by the two parties, to the State Legislature
01/10/2022	The State Legislature rejected both plans. The law required the IRC to submit a new plan within 15 days.
01/24/2022	The IRC decided not to propose a new plan as it was deadlocked.
02/03/2022	The State Legislature, controlled by the Democratic Party, then passed its own plans and the governor, also a Democratic, signed it into law.
04/21/2022	In the ruling of the lawsuits against the district maps approved by the governor, the State of New York Court of Appeals struck down the Congress and State Senate districting maps.
05/16/2022	The special master appointed to redraw New York's legislative districts by the court released the draft maps
05/21/2022	The court released the final maps for New York's 26 congressional and 63 state Senate districts. The State Legislature did not challenge or amend the maps.

The primary objective of this study is to use quantitative metrics to evaluate the degree of gerrymandering in the four sets of congressional district maps for New York State proposed

by the IRC, the State Legislature, and the State Court of Appeals, respectively. Unlike most metrics, the study employs a metric with more comprehensive criteria to assess districts' compactness and gerrymandering, taking into account factors such as population distribution, state boundaries, and geospatial topology.

2. Data and Methods

To conduct the evaluation, the census population data and four sets of congressional district maps were collected from public sources (Table 2).

Table 2. Data for Assessing Gerrymandering of District Maps in New York State

Data	Source	URL	Format
Census Population Enumeration	Decennial Census P.L. 94-171 Redistricting Data	https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html	Boundary data in Shapefile; Population data in binary format (with import script code)
District Maps submitted by IRC	New York State IRC Plans 2021/2022	https://www.nyirc.gov/plans	Spatial/GIS Data in Shapefile
District Maps by the Legislature and Court	NYS Legislative Task Force on Democratic Research and Reapportionment	https://latfor.state.ny.us/maps/ Also historial archive of the website at https://archive.org/web/	Spatial/GIS Data in Shapefile; PDF maps

While it is straightforward to collect and process the data for gerrymandering assessment, it is rather challenging to quantitively identify and measure gerrymandering despite the availability of numerous metrics. Of the two main categories of gerrymandering metrics or tests, one is based on the deviation of the election results from those implied by the popular vote. One notable example is the efficiency gap (Bernstein & Duchin, 2017; Stephanopoulos & McGhee, 2015). However, these election results-derived metrics essentially converge to the popular vote and ultimately suggest the invalidity of the current electoral system. The other category directly measures the compactness of the boundaries and can be applied without election information (Young, 1988). Existing compactness measurements for gerrymandering, however, target at particular aspects of gerrymandered shapes such as elongation, indentation, bizarreness, or dispersion without adequately integrating them with

spatial context (Fan, Li, Wolf, & Myint, 2015; Lunday, 2014; MacEachren, 1985). Gerrymandered shapes are geometrically complex with multidimensional characteristics, yet most of those geometry-based metrics can only address one aspect and fail to consider the geographic context such as population and sub-population distribution, external boundary constraints, and internal topology (Chambers & Miller, 2013; Niemi, Grofman, Carlucci, & Hofeller, 1990). Most significantly, they only rank districts without offering a cut-off value to consistently identify gerrymandered boundaries.

This paper employs a quantitative gerrymandering metric based on non-overlapping maximum coverage circles that is proposed by Sun (2021). This metric comprehensively and coherently integrates population, boundary constraints, and spatial context. It also reflects roundness, convexity, and closeness. Most noticeably, it offers a natural threshold of zero for gerrymandering identification, which can conservatively but directly and unambiguously identify gerrymandered boundaries. In addition to this comprehensive metric, other simpler measures are also calculated for the purpose of comparison. All the measures for the four sets of electoral district maps are compared statistically.

3. Results

The 2020 census has led to a reduction in the number of seats allocated to the State of New York in the House of Representatives, from 27 to 26. Consequently, there is a need to redraw the congressional districts, which will certainly result in the displacement of at least one incumbent member of the House.

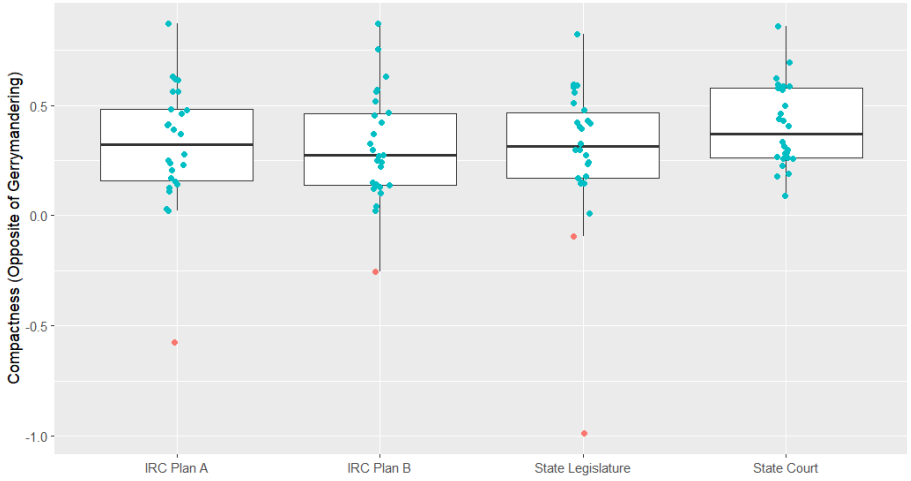


Figure 1. Gerrymandering Measurements of Congressional Maps Proposed for 2022 Mid-Term in New York State. The metric is based on coverage circle path distances proposed by Sun (2021). Negative values (red) indicate clear gerrymanders.

The new congressional district maps submitted by the IRC—including Plan A and Plan B, the State Legislature, and the State Court, as well as the map for the previous decade, are available as interactive web apps at *https://suncodeearth.github.io/nys_cd_maps*. The actual map used for the election is the one proposed by the Court.

From the quantitative measures using the maximum coverage circle path distance-based metric (CCPD) and others, it is clear that the Court-drawn map has higher average and median compactness than those proposed by the IRC and the legislature (Figure 1, Table 3). Note that Moment of Inertia (MOI) related metrics are measuring dispersion, which is positively correlated to gerrymandering. Other metrics, including CCPD, measure compactness, which is negatively correlated to gerrymandering. With the Kruskal-Wallis rank sum test, the IRC Plans are not statistically different (p^1>0.05) from the one proposed by the State Legislature on multiple gerrymandering metrics with or without the consideration of population, state boundary, or spatial topology (Table 3). To the contrast, the difference between the State Court-drawn map is statistically different from other maps, particularly when measured without population (p^2<0.05). This also implies that the Court-drawn districts appear much more compact, although they still bear much gerrymandering when population distribution is considered.

Table 3. Measures of Compactness and Gerrymandering of District Maps in New York State

Metric	IRC Plan A*	IRC Plan B	State Legislature	p^1	State Court	p^2
Polsby Popper	0.28 (0.12) [0.10, 0.60]	0.28 (0.13) [0.13, 0.58]	0.25 (0.09) [0.05, 0.43]	0.9	0.36 (0.11) [0.18, 0.61]	<0.01
Moment of Inertia (MOI) **	1.79 (0.84) [1.07, 5.24]	1.80 (0.67) [1.13, 4.36]	2.02 (1.07) [1.13, 6.49]	0.5	1.54 (0.37) [1.04, 2.41]	0.027
Population weighted MOI **	1.57 (1.17) [0.27, 6.37]	1.59 (0.96) [0.27, 5.15]	1.75 (1.43) [0.33, 7.97]	0.8	1.22 (0.49) [0.23, 2.20]	0.13
Coverage- Circle Path Distance (CCPD)	0.27 (0.22) [-0.33, 0.67]	0.26 (0.18) [-0.12, 0.61]	0.21 (0.24) [-0.62, 0.56]	0.7	0.35 (0.16) [0.07, 0.67]	0.042
Population Weighted CCPD	0.32 (0.28) [-0.58, 0.87]	0.31 (0.25) [-0.26, 0.87]	0.29 (0.33) [-0.99, 0.82]	>0.9	0.41 (0.19) [0.09, 0.86]	0.2

* *Mean (SD) [Min, Max].*
** *MOIs measure dispersion, one characteristic of gerrymandering; other metrics are indicators of compactness, the opposite of gerrymandering.*
[1] *Kruskal-Wallis rank sum test for the differences among IRC Plan A & B, and State Legislature*
[2] *Kruskal-Wallis rank sum test for the difference between State Legislature and State Court*

In addition, the map drawn by the Court also has much less extremely gerrymandered shapes. Noticeably, the map drawn by the State Legislature has two clear cases of gerrymandering, while the IRC plans have one for each. The politicians at the State Capital did not eliminate gerrymandering; instead, they made it even worse. The independent special master appointed by the court, on the contrary, divided that district and made it more compact (Figure 2).

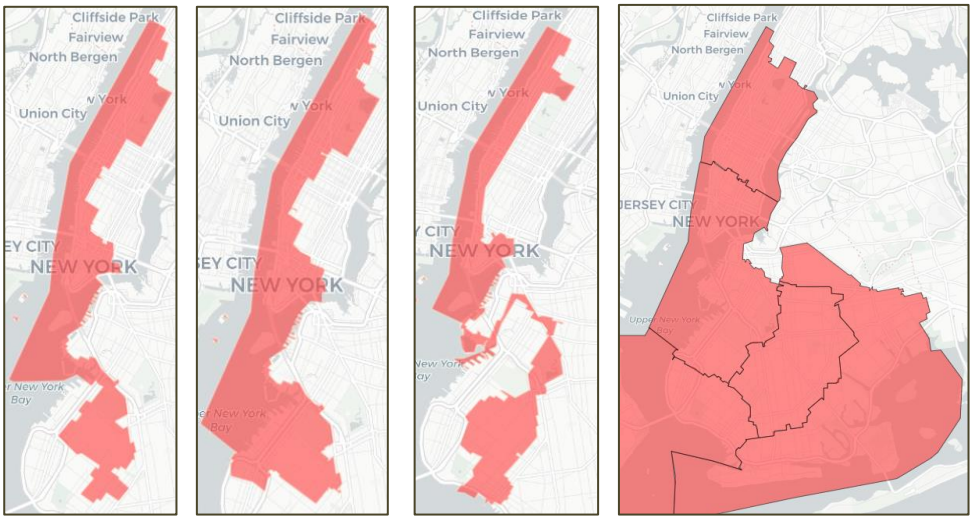


Figure 2. The most gerrymandered district in the plans proposed by IRC Plan A, IRC Plan B, and the State Legislature was revised in the Court-drawn map (from left to right).

4. Conclusion

This paper evaluates the degree of and identifies the cases of gerrymandering in four different sets of congressional district maps for the 2022 mid-term elections in New York State. There are two important implications from the study. First, it is critical to establish a legally and politically accepted metric or test for the identification of extreme gerrymandering. Due to the nature of redistricting, some wiggle room for political flexibility is practically necessary and favorable, which implies that gerrymandering cannot be completely avoided. However, identifying extreme gerrymandering with a quantitative metric helps avoid proposing or submitting district maps that would be rejected by the legislature or challenged in the courts. This metric must have a clearly defined and undisputable cut-off value. The maximum coverage circle path distance metric used in this paper seems appropriate for this task and warrants more case studies. Second, with the polarization of the American voters and politicians, relying on the political system to produce new electoral maps seems inefficient and problematic. Deeply influenced by the results of redistricting, it is impossible for those in the political system to withhold their strongly motivated influences, either directly or

through "independent" commissions. This paper clearly shows that the special master appointed by the court, without direct connection to the State Legislature or other political groups, produced the least gerrymandered congressional district map. For the 2012 election, it was also the court that drew the final congressional district map. Considering the fact that increasingly more redistricting cases end up in court, it might be the time to consider permanently shifting the power of drawing electoral maps to the legal system.

References

Abramowitz, A. I., Alexander, B., & Gunning, M. (2006). Incumbency, redistricting, and the decline of competition in US House elections. *The Journal of Politics, 68*(1), 75-88.

Ansolabehere, S., & Snyder Jr, J. M. (2012). The effects of redistricting on incumbents. *Election Law Journal, 11*(4), 490-502.

Bernstein, M., & Duchin, M. (2017). A Formula Goes to Court: Partisan Gerrymandering and the Efficiency Gap. *Notices of the American Mathematical Society, 64*(09), 1020-1024. doi:10.1090/noti1573

Chambers, C. P., & Miller, A. D. (2013). Measuring legislative boundaries. *Mathematical Social Sciences, 66*(3), 268-275. doi:10.1016/j.mathsocsci.2013.06.001

Crocker, R. (2012). *Congressional Redistricting: An Overview*. Retrieved from Washington DC:

Fan, C., Li, W., Wolf, L. J., & Myint, S. W. (2015). A Spatiotemporal Compactness Pattern Analysis of Congressional Districts to Assess Partisan Gerrymandering: A Case Study with California and North Carolina. *Annals of the Association of American Geographers, 105*(4), 736-753. doi:10.1080/00045608.2015.1039109

Lunday, B. J. (2014). A metric to identify gerrymandering. *International Journal of Society Systems Science, 6*(3), 285-304.

MacEachren, A. M. (1985). Compactness of geographic shape: Comparison and evaluation of measures. *Geografiska Annaler: Series B, Human Geography, 67*(1), 53-67.

Niemi, R. G., Grofman, B., Carlucci, C., & Hofeller, T. (1990). Measuring Compactness and the Role of a Compactness Standard in a Test for Partisan and Racial Gerrymandering. *The Journal of Politics, 52*(4), 1155-1181. doi:10.2307/2131686

Stephanopoulos, N. O., & McGhee, E. M. (2015). Partisan gerrymandering and the efficiency gap. *The University of Chicago Law Review, 82*(2), 831-900.

Sun, S. (2021). Developing a Comprehensive and Coherent Shape Compactness Metric for Gerrymandering. *Annals of the American Association of Geographers, 111*(1), 175-195. doi:10.1080/24694452.2020.1760779

Young, H. P. (1988). Measuring the compactness of legislative districts. *Legislative Studies Quarterly, 13*(1), 105-115.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16039

A Machine Learning approach to constructing weekly GDP tracker using Google Trends

Jean Christine A. Armas[1], Cherrie R. Mapa[1], Ma. Ellysah Joy T. Guliman[1], Michael Lawrence G. Castañares[1], Genna Paola S. Centeno[1]
[1]Department of Economic Research, Bangko Sentral ng Pilipinas, Philippines

Abstract
The outbreak of the COVID-19 pandemic further highlighted the limitation of existing traditional indicators as policy formulation, particularly during crisis periods, demands timely and granular data. We construct the first Weekly Growth Domestic Product (GDP) Tracker in the Philippines using topic- and category- based Google Trends search volumes with the aid of machine learning models. We find that our Weekly GDP Tracker is a useful high-frequency tool in nowcasting economic activity. We also show that the machine learning-based GDP tracker outperforms the traditional autoregression models under study in terms of lower root mean square error (RMSE) for both train and test datasets. On the whole, our Weekly GDP Tracker can serve as a useful complementary surveillance tool for monitoring economic activity.

Keywords: *Nowcasting; GDP; Google Trends; machine learning models; neural networks.*

1. Introduction

Timely and accurate information are essential inputs for policy formulation. Data becomes even more important during crisis periods, such as the COVID-19 pandemic, as high frequency and granular data are integral in crafting prompt and appropriate policy responses that can help attenuate the impact of a crisis. However, official economic statistics are typically published with a significant time lag. In the case of the COVID-19 pandemic, an extra layer of challenge emerged as collection of official statistics was hampered by the imposition of mobility restrictions during the height of the health crisis. These motivate the interest of policymakers, including monetary authorities, to tap alternative data sources to supplement existing official statistics or traditional indicators.

In the area of macroeconomic surveillance, the Gross Domestic Product (GDP) is the official and most comprehensive indicator for measuring economic activity. In the Philippines, the GDP is available on a quarterly basis and is published by the Philippine Statistics Authority (PSA) 40 days after the reference quarter except for the Q4 GDP which is released after 30 days.[1] Due to this publication lag, the Bangko Sentral ng Pilipinas (BSP) uses a number of models to nowcast the GDP as information on the output growth and the cyclical position of the economy in the business cycle are important considerations in policy formulation. Thus far, the BSP's nowcasting models for the GDP have employed traditional statistics with nowcast updates implemented on a monthly or quarterly basis.

This study attempts to build a high-frequency indicator of GDP growth that capitalizes on the use of alternative data. Our objectives are: (a) to construct a weekly GDP Tracker that can nowcast quarterly GDP growth using machine learning models trained on pre-identified Google search volumes; and (b) to evaluate the usefulness of Google Trends data in nowcasting GDP pending the availability of official statistics. To ensure that the use of Google Trends data is suitable for statistical and economic analysis, the topic- and category-based searches are selected based on the authors' expert and sensible judgment on mapping relevant internet searches with the components of National Accounts of the Philippines (NAP). To the authors' knowledge, this is the first empirical research in the Philippines to use both the topic- and category- based Google Trend searches in nowcasting GDP.

2. Data: Google Trends Data Selection and Statistical Pre-processing

This study uses data from two sources: PSA and Google Trends. The quarterly real GDP data (with 2018 base year) are sourced from the PSA. No other statistical processing was

1 PSA, Technical Notes on the National Accounts of the Philippines, available online at https://psa.gov.ph/statistics/technical-notes/node/168102.

performed for the GDP data apart from taking the difference in the natural logarithm of the real GDP to derive the year-on-year GDP growth rates. Thus, this section delves into the selection of Google Trends indicators incorporated in our study and the statistical pre-processing steps applied to these indicators.

2.1. Google Trends Selection

Google Trends data is an analytical tool provided by Google that allow users to determine relative interest on a particular search term. It shows how frequently a search term is entered into Google's search engine relative to all Google searches for a particular geographical region and time. This study utilizes Google search volume indices for the main categories, selected sub-categories and authors' pre-identified topics that are relevant to the estimation of GDP growth (Table 1).

Table 1. List of Select Google Trends Variables

a. Google Trends categories and sub-categories		
Arts & Entertainment	Health	Pets & Animals
Autos & Vehicles	Hobbies & Leisure	Real Estate
Beauty & Fitness	Home & Garden	Reference
Books & Literature	Internet & Telecom	Science
Business & Industrial	Jobs & Education	Shopping
Computers & Electronics	Law & Government	Sports
Finance	News	
Food & Drink	Online Communities	
Games	People & Society	
Advertising & Marketing	Chemicals Industry	Pharmaceuticals & Biotech
Aerospace & Defense	Construction & Maintenance	Printing & Publishing
Agriculture & Forestry	Energy & Utilities	Retail Trade
Automotive Industry	Enterprise Technology	Small Business
Business Education	Entertainment Industry	Textiles & Nonwovens
Business Finance	Hospitality Industry	Transportation & Logistics
Business News	Industrial Materials & Equipment	
Business Operations	Manufacturing	
Business Services	Metals & Mining	
b. Google Trends topics		
Job	Subsidy	Unemployment Benefits
Pantawid Pamilyang Pilipino Program	Tax	Investment
	Unemployment	Resignation

Source: Google Trends; Authors' Selection

2.2. Data Pre-processing

Google Trends data series are collected on monthly (January 2004 - August 2022) and weekly (January 2014- August 2022) frequencies. For the monthly data series, corrections for known breaks in the time series data are implemented.

For the period 2004 to 2022, Google reported the following three (3) breaks in their time series data: (1) January 2011 due to geographical localization; as well as (2) January 2016 and (3) January 2022, both due to improvements in the data collection system. Only the breaks in 2011 and 2016 are corrected to address the potential issue of spikes in growth rates that may be attributed to changes in the data collection methods of Google.[2]

The breaks are addressed by introducing an adjustment that results into zero growth rate at the breakpoint. A backward correction approach was used to correct for the breaks starting from January 2016 back to January 2011. This approach, however, is a deviation from the study of Woloszko (2020) that made use of forward correction to address the breaks.[3]

3. Machine Learning Models

This study takes advantage of the empirically-documented ability of machine learning algorithms to generate relatively accurate and robust predictions. Machine learning models offer two key advantages: (a) capture the non-linearities in the data that could better explain the movements in real GDP especially during periods of extreme economic stress and heightened uncertainty; and (b) handle a wide array of variables without running into issues of overfitting through the multi-layer structure of machine learning models.

In this study, four (4) machine learning techniques are evaluated namely, Support Vector Regression (SVR), Decision Trees, Random Forest, and Artificial Neural Network (ANN). For all these models, the data are split into a train dataset (first 65 quarters) and a test dataset (last 5 quarters). Each model is trained to learn the patterns from the train dataset. Model performance in both the training and test sets are evaluated by computing for the root mean square error (RMSE). The results are also compared with the traditional autoregressive models. Table 2 below summarizes model performance based on RMSE for the train and test datasets. Notably, the ANN has outperformed other machine learning models and even the

[2] The impact of the January 2022 break on the data series is not yet evident at the time of the model construction.

[3] The backward break correction is done on each variable starting at the earliest date (i.e., 2004). For each variable, the difference between January 2011 (2016) and January 2010 (2015) is added to observations earlier than January 2011 (2016) inclusive. The break in 2022 will be addressed once there is evidence of a significant shift in the time series data.

traditional autoregression models. Therefore, it was chosen to be the main model for this study.

Table 2. Forecast evaluation of machine learning models vs. traditional time series models

	RMSE using Train data	RMSE using Test data
SVR	2.55	2.29
Decision Trees	0.50	12.41
Random Forest	1.04	5.62
ANN	0.53	1.49
ARIMA (1,1,1)	2.83	7.26
AR(1)	2.67	6.33

Source: Authors' estimates

4. Construction of the Weekly GDP Tracker

The Weekly GDP Tracker capitalizes on the use of a more frequent Google Trends data series to be able to extract leading information from this type of unconventional data source and thus, infer sensible predictions on economic activity or variations in business cycle especially during unprecedented periods.

The Tracker is constructed using a two-step model approach, broadly following Woloszko's (2020) methodology. First, the machine learner is trained to predict the quarterly GDP growth using topic- and category- based Google Trends searches, as outlined in the previous section. Second, by employing the frequency-neutrality assumption in Woloszko (2020), the estimated elasticities from the quarterly model were applied to the weekly Google Trends data series.

The application of elasticities from quarterly model on the weekly Google Trends requires calibrating the weekly Google Trends series to match the beak-corrected monthly Google Trends indices and using its 12-week moving average as input in the trained model.

5. Analysis of the Model Results

An important question on the use of the Weekly GDP Tracker is how well it can nowcast economic activity? Overall, the results show that the Weekly Tracker is a useful complementary surveillance tool to official statistics in terms of providing relevant leading information on the likely trend or path of output growth as well as capturing the important crisis episodes or business cycle fluctuations.

We find relatively good prediction performance of our chosen machine learning model, primarily due to its ability to find the optimal parameters that could fit the training dataset. At the same time, as presented in Table 2, the RMSE for both training and test datasets

percent of the ANN are significantly lower when compared to the traditional time series models such as Autoregression (AR) and Autoregressive Integrated Moving Average (ARIMA). The predictive performance of our quarterly model is also assessed with regard to the movements of actual output growth during the 2020 COVID-19 pandemic. We note that the COVID-19 crisis period as well as the subsequent rebound following the downturn are well-captured by the predicted values for output growth. In particular, our Weekly GDP Tracker was able to capture about 96 percent of the slump observed in actual GDP growth in Q2 2020.

The weekly nowcast GDP growth from January 2015 to August 2022 is plotted against the actual quarterly year-on-year GDP growth in Figure 1. As shown, the Weekly Tracker closely tracks the general direction of the actual GDP, pointing to the usefulness of the Weekly Tracker in providing reasonable near real-time measurement of economic activity.

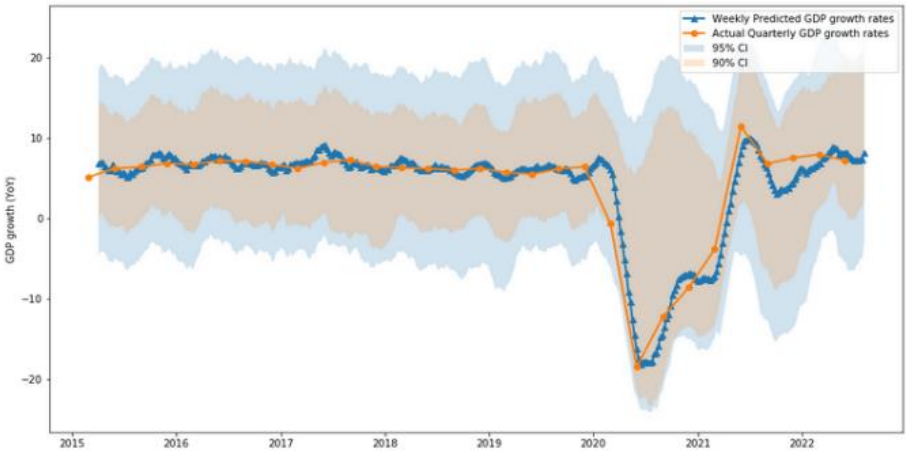


Figure 1. Weekly GDP tracker and actual quarterly GDP (in percent, %) (January 2015 – August 2022). Source: Authors' estimates

5.1. Most Important Predictors of GDP Growth Based on SHAP Values

One of the common problems with machine learning algorithms, especially with neural networks, is their black-box nature that makes analytical interpretation of the results difficult. One way to address this is to use an interpretability technique known as the Shapley values, which estimate the average marginal contribution of a variable to the prediction over all possible variable combinations or coalitions. This study uses the SHapley Additive exPlanations (SHAP) a fast algorithm implementation to extract Shapley values.

Based on the train dataset, Google searches for investment, unemployment, real estate, business news, and agriculture and forestry along with consumption-related searches for autos and vehicles and hobbies and leisure are found as the top contributors to the predicted GDP growth rate based on their SHAP values (Figure 4). Intuitively, higher google searches for investment and real estate may be correlated with higher economic growth while higher searches for unemployment may be associated with weaker GDP growth.

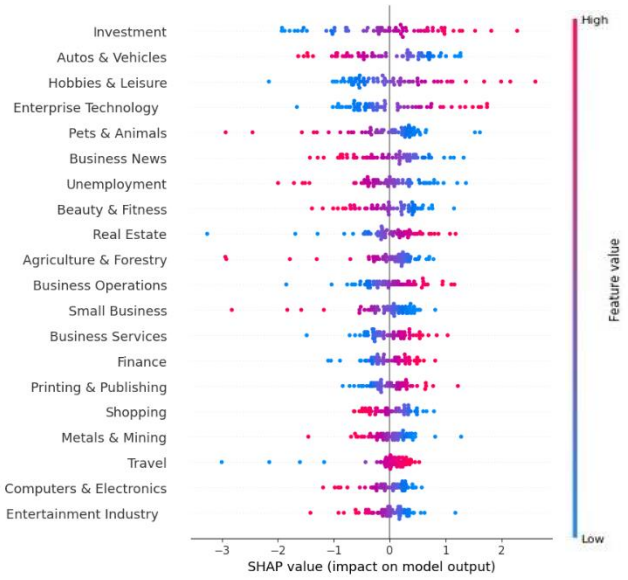


Figure 2. Most important predictors of GDP based on SHAP value. Source: Authors' estimates Note: The color bar in the illustration corresponds to the raw values of the variables for each instance on the graph. That is, variables with high and low values appear as red and blue dots, respectively. Meanwhile, the horizontal axis shows whether the effect of that value is associated with a higher or lower prediction. For example, high searches for investment (denoted by red dots) have positive impact on predicted GDP growth (shown in the horizontal axis).

6. Conclusion

The pandemic highlighted the importance of unconventional data sources, such as internet data, in macroeconomic surveillance. To the authors' knowledge, this is the first empirical research in Philippine literature to capitalize on the use of Google – the world's largest search engine– in near-real time tracking of output growth. To ensure that the use of Google Trends is suitable for statistical and economic analysis, topic- and category- based searches are selected based on the authors' expert and sensible judgment.

In this study, the Weekly GDP Tracker was constructed broadly following the approach of Woloszko (2020). The evaluation shows that the Weekly GDP Tracker is a useful high-

frequency tool in tracking economic activity. The broad goal is not to replace the existing suite of economic models by the BSP but to contribute to surveillance by providing a high frequency monitoring tool that takes advantage of the readily available alternative data and the predictive capacity of advanced algorithms. Thus, pending the availability of quarterly national accounts, the Weekly GDP Tracker can serve as complementary surveillance tool of economic activity.

Like all models, prediction errors are anticipated, especially when the sample size to train or learn from is limited. Nonetheless, a key advantage of the weekly tracker is its greater flexibility to re-train and update nowcast predictions for GDP as new actual data becomes available. Hence, the model's learning capacity to predict output growth estimates can be further improved as new data comes in.

References

Armas, J. and P. Tuazon (2021). 'Revealing investors' sentiment amid Covid-19: the Big Data evidence based on internet searches' in Bank for International Settlements eds., New developments in central bank statistics around the world, vol. 55, Bank for International Settlements.

Assunção, J., Fernandes, P. (2022). 'Nowcasting GDP: An Application to Portugal'. Forecasting. 4, pp. 717-731, https://doi.org/10.3390/forecast4030039.

D'Amuri, F. and J. Marcucci (2012). 'The predictive power of google searches in forecasting unemployment.'. Bank of Italy Working Paper 891.

Gil, M., J. Perez, A. Sánchez, and A. Urtasun (2018): "Nowcasting private consumption: traditional indicators uncertainty measures, credit cards and some internet data". Working Paper 1842, Banco de España.

Haykin, S. (2009) Neural Networks and Learning Machines, 3rd edition, Pearson Prentice Hall.

Indaco, A. (2020). From twitter to GDP: Estimating economic activity from social media. Regional Science and Urban Economics.

Molnar, C. (2020), Interpretable Machine Learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/.

Shapley, Lloyd S. (1953). "A value for n-person games." Contributions to the Theory of Games 2.28 pp. 307-317

Tuhkuri, J. (2016). 'Forecasting Unemployment with Google Searches'. ETLA Working Papers, No. 35 The Research Institute of the Finnish Economy (ETLA), Helsinki

Vosen, S. and Schmidt, T. (2011). 'Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends'. Journal of Forecasting. 30, pp. 565-578. DOI: 10.1002/for.1213

Woloszko, N. (2020), "Tracking activity in real time with Google Trends", OECD Economics Department Working Papers, No. 1634, OECD Publishing, Paris, https://dx.doi.org/10.1787/6b9c7518-en

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16491

An estimate of the Italian Consumer Confidence Index at regional level using Google Trends data

Josep Domenech[1], Andrea Marletta[2]
[1]Department of Economics and Social Sciences, Universitat Politècnica de València, Spain
[2]Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy.

Abstract

Data about consumer confidence indices are often used as a gauge of the entire economy of a country. In Italy, this information is collected by Istat and it is available at national level and at the first sub-level, the geographic area, but not at the regional level. Previous research has demonstrated that the volume of some Google searches are correlated with the consumer confidence. Since Google Trends data are available both at national and regional level, the aim of this paper is to explore they can be combined with the data offered by Istat to obtain an estimate for consumer confidence indices at the second sub-level, i.e., the regional area. To this end, a set of search topics and words have been selected as potential predictors to acquire more information about consumer confidence indices for 20 Italian regions from 2007 to 2022. The obtained regional estimates are in line with the geographic area being successful to identify the periods of economic crisis due to the 2008 financial crisis and the 2020 Covid-19 pandemic.

Keywords: *Consumer confidence, Google Trends, Non-traditional data sources*

1. Introduction

The Consumer Confidence Index (CCI, henceforth) is a very important indicator to measure the health condition of the economy of a country. Many contributions showed the strict link between these indices and the economic activity (Golinelli and Parigi, 2004; Kilic and Cankaya, 2016). Other contributions focused the attention on the predictive power of these meaures to forecast the consumption spending (Malgarini and Margani, 2007; Dees and Brinca, 2013).

For these reasons for all the countries in the European Union, the survey on consumer confidence is part of the joint harmonized EU program of business and consumer surveys. In Italy, this survey is realized monthly by the Italian National Institute of Statistics (ISTAT) with the CATI (Computer Assisted Telephone Interviewing) technique. The observed phenomena are consumer assessments and expectations and the final data dissemination is realized producing a final indicator called CCI and a group of 4 sub-indicators measuring the personal, economic, current and future confidence of the interviewers. Finally, these indicators are published as index numbers (reference year 2010 =100).

From a geographic point of view, the CCI is available at national level and at the first sub-level, a modified version of NUTS-1 (Nomenclature of Territorial Units for Statistics). Usually, in Italy, the first level of NUTS involves 5 areas, North-West, North-East, Centre, South and Islands. For CCI, data are available in 4 areas: North-West, North-East, Centre and South (also including Islands). No information is at NUTS-2 level, that is to say, the 20 single regions.

An index about regional consumer confidence is missing and the aim of this paper is to verify whether the existence of Google Trends (GT) time-series about some economic terms related to this issue could help in the estimation of such index. The birth of GT data in 2004 gave the start to a literature in combining these data with macro-economic variables mixing official and non-traditional data sources. GT data have often been associated to predictions for unemployment rates at national level (Naccarato et al., 2018; Simionescu and Cifuentes-Faura, 2021), other authors underlined the attention on smallest area at regional level (Falorsi et al., 2017; Bartha and Bontempi, 2022). Finally, some papers tried to explain the relation between economic indicators and GT data during Covid-19 pandemic (Lee, 2020; Simionescu and Raisiene, 2021).

The availability of GT time-series on real time at regional level for Italy could be exploited to control what is the relationship between official estimates for CCI at NUTS-1 level and the searches in Google for terms and topics related to CCI at NUTS-2 level.

The paper is organized as follows: after the introduction, Section 2 is devoted to the methodology and the data collection, in Section 3 some preliminary results have been presented and finally in Section 4 some conclusions will follow.

2. Methods and data collection

The theoretical background of the methodology is related to the conjoint use of data from different sources, from a statistical point of view it involves a linear regression based on some techniques of multivariate analysis. In the linear regression model, the dependent variable is the estimate for regional CCI and the independent variables are the results of the multivariate analysis. To extract the similarities between official and GT data, the following process is described. The procedure for computing a regional estimate of CCI could be resumed in 4 steps, similar to the process described by Eichenauer et al. (2022) and Woo and Owen (2019).

Firstly, the search volumes for some terms potentially correlated with the CCI are retrieved and seasonally adjusted; secondly, among all terms, only those highly correlated with the CCI are selected; thirdly, a Principal Component Analysis (PCA) of the selected series is conducted to extract the common signal in them (Jolliffe, 2002); finally, the first Principal Component is used to estimate the parameters of a linear regression model on the CCI.

After the model for the NUTS-1 level CCI is estimated, the search volumes at the regional level for the same terms are retrieved from GT seasonally adjusted. Using the PCA factor loadings of the national data, factors at the regional level can be computed. Then, using the model parameters estimated, it is possible to estimate the CCI at a regional level.

The first principal component was used to estimate:

$$CCI_t = \beta_0 + \beta_1 GT_t + u_t$$

where CCI is the Consumer Confidence Index at Nuts-1 level and GT is the first principal component extracting the common signal of the selected search term volumes.

The method described above has been applied to the CCI in Italy. A list of terms, comprising keywords and topics related to the economic sentiment, was requested to GT. Those with a Pearson's correlation coefficient higher than 0.7 were used as input for the PCA.

Data for CCI and GT time-series are referred to the period from 2007 to 2022. Since the volatility of GT data, multiple extractions have been achieved in January and February and the final result is the average time series of multiple extractions.

GT data have been collected using the library *gtrendsR* and *trendecon* in the statistical software R. The dataset referred to 33 searches (15 topics and 18 terms) related to the domain of the macroeconomic phenomena as the CCI. The list of the terms is the following: “production”, “unemployment”, “employment”, ”labour”, ”recession", "inflation", ”public debt”, “economic crisis”, “recovery”, “minimum income guaranteed”, ”unemployment allowance”, “failure”, ”purchase”, “saving”, ”bankrupt”, “mortgage”, ”sales”, ”construction”. The time span covered from January 2007 to December 2022 using monthly data for 21 geographical areas, the national searches and the 20 regional searches. Data about CCI have been achieved using the Istat datawarehouse available at “dati.istat.it”. This dataset is available as monthly data from January 2005 to December 2022 for 5 geographical areas, the national level and the 4 sub-national levels (North-West, North-East, Centre and South (also including Islands).

3. Results

In this section, preliminary results are shown for Central Italy and its regions coloured in Figure 1. The geographic area named Central Italy is composed of 4 regions: Tuscany, Umbria, Marche and Lazio. From a geographical point of view, it represents about 20% of the total population of Italy and 19% of the total area of Italy. The most representative cities of this area are Rome in Lazio and Florence in Tuscany. From an economic point of view, this area is characterized by a strong industrial activity realized by small and medium enterprises and by an important touristic sector.

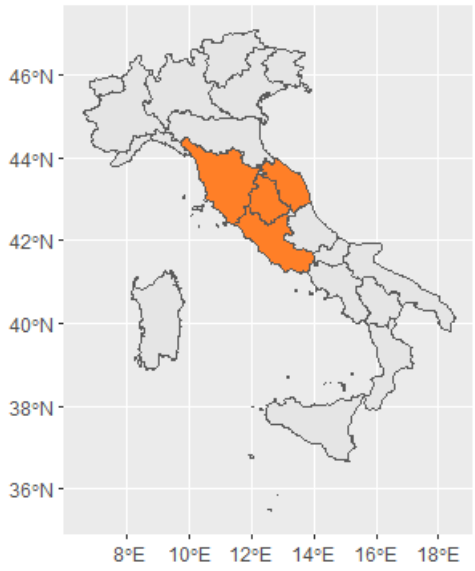


Figure 1. Map of Italy and Central Italy (coloured).

This area has been selected as a referring point for this work because it is probably the most homogenous area able to better represent the entire Italian territory. This statement is confirmed by the Figure 2, in which CCI time series from 2007 to 2022 is displayed for Italy and Central Italy. It is possible to note that the perception is very similar with a correlation coefficient equal to 0.98, this means that in Central Italy the consumer confidence is essentially very close to the national one. For the data collection process of Istat, no information is available about how this confidence is percieved in the 4 regions and the motivation of this work is trying to give a representation of CCI at regional level using an alternative data source.

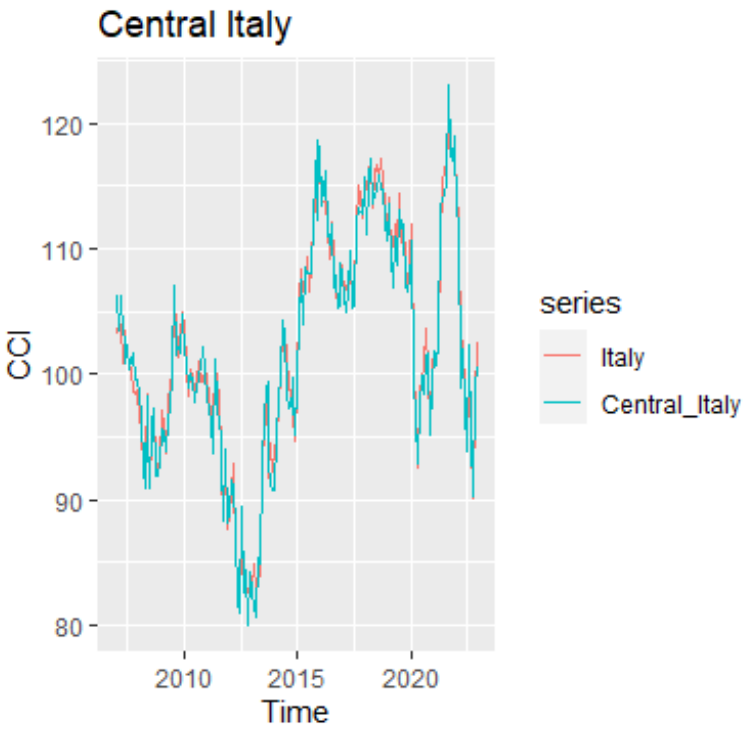


Figure 2. CCI for Italy and Central Italy (2007-2022). Source:Istat

Using the methods described in the previous section, an estimate for CCI based on GT data is displayed in Figure 3. Among all the terms used, the terms with the higher correlation with the CCI in Central Italy are: economic crisis (topic), sales (topic), unemployment allowance (topic), flights (topic), economic crisis (term).

These preliminary results are satisfactory because for Tuscany, Marche and Lazio the correlation coefficient between the regional estiamate and the official CCI for Italy and Central Italy is higher than 0.5 (see Table 1).

Table 1. Correlation between regional CCI and Central Italy (2007-2022). Source: Istat and GT data

Region (Central Italy)	*Correlation with Central Italy*	*Correlation with Italy*
Tuscany	0.551	0.566
Umbria	0.448	0.464
Marche	0.543	0.559
Lazio	0.581	0.597

Probably, results for Umbria are less stable, because it is the smallest region of the Central area and GT data for very small areas are not reliable.

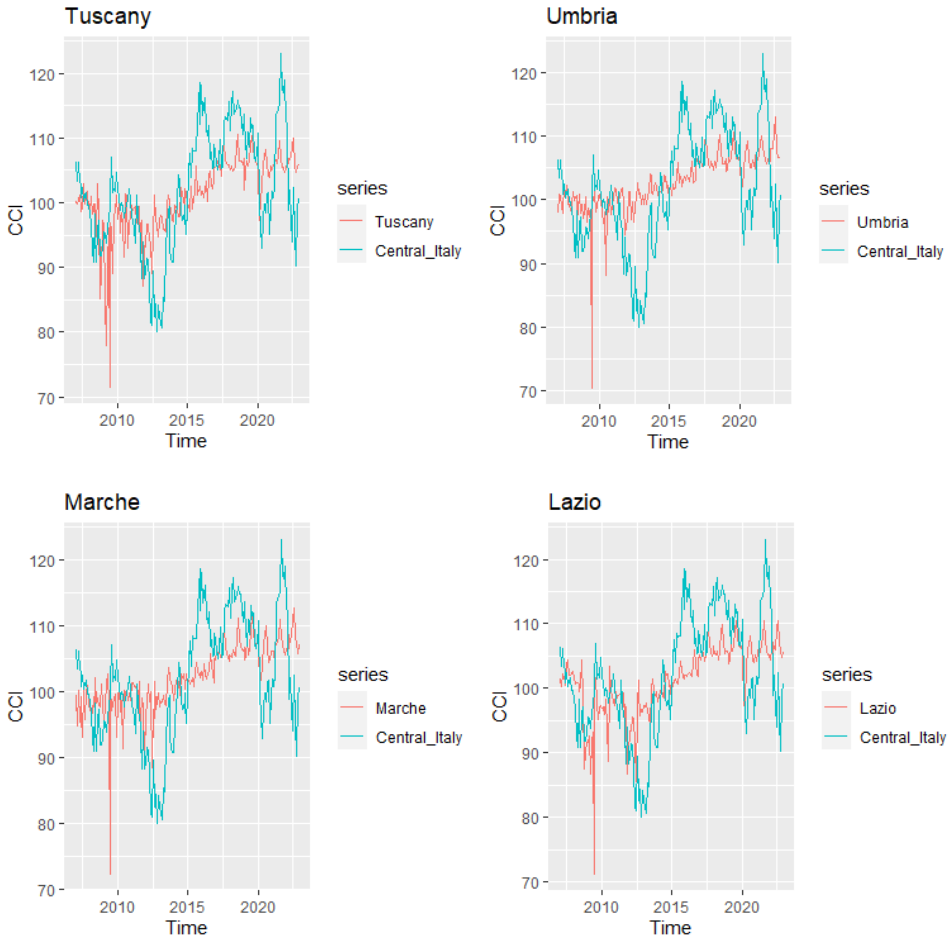


Figure 3. Estimated regional CCI compared to Central Italy (2007-2022). Source:Istat and GT data.

From a graphical point of view, all GT time series have a fall in 2009, probably due to the financial crisis. With respect to the decrease of CCI in 2012 due to the spread crisis, according to GT data, it is well recognized in Lazio. After this period, there is a slight increase for all the regions without the peak of 2016 and 2018 amounted by the real data. Relatively to the Covid-19 pandemic, the fall of the consumer confidence is present in Tuscany and Lazio.

4. Conclusions

In this work an approach to obtain an estimate for Consumer Confidence Index at regional level has been presented for Italy from 2007 to 2022 using Google Trends data. Istat provided monthly data for CCI at national and NUTS-1 level but not at regional level. On the other hand, GT data have been collected using some terms and categories related to the consumer confidence for the same period at national and regional level. Combining these two official and unofficial sources, and using a simple approach based on principal component analysis, a first estimate for CCI have been achieved for the 4 regions of Central Italy. The regional estimates are affected by the typical volatility of GT data, but especially for Lazio and Tuscany, the regional CCI presents good values of correlation with official CCI for Central Italy and a fair capability to catch sudden drops due to some crisis.

Some future works could regard the introduction of some error measurements to evaluate the goodness of the model, or a different procedure to select the terms to insert in the estimate model. Finally, some enhancements could involve the integration of some prediction methods for the time-series analysis of the estimated values.

Acknowledgments

This work was partially funded by MCIN/AEI/10.13039/501100011033 under grant PID2019-107765RB-I00.

References

Bartha, S., & Bontempi, M. E. (2022). Measuring Economic Uncertainty for Poland.

Dees, S., & Brinca, P. S. (2013). Consumer confidence as a predictor of consumption spending: Evidence for the United States and the Euro area. *International Economics*, *134*, 1-14.

Eichenauer, V. Z., Indergand, R., Martínez, I. Z., & Sax, C. (2022). Obtaining consistent time series from Google Trends. Economic Inquiry, 60(2), 694-705.

Falorsi, S., Fasulo, A., Naccarato, A., & Pratesi, M. (2017, July). Small area model for Italian regional monthly estimates of young unemployed using Google trends data. In *61st World Congress of the International Statistical Institute* (pp. 16-21).

Golinelli, R., & Parigi, G. (2004). Consumer sentiment and economic activity: a cross country comparison. Journal of Business Cycle Measurement and Analysis, 2004(2), 147-170.

Jolliffe, I.T., (2002). Principal Component Analysis, second edition, New York: Springer-Verlag New York, Inc.

Kilic, E., & Cankaya, S. (2016). Consumer confidence and economic activity: a factor augmented VAR approach. Applied Economics, 48(32), 3062-3080.

Lee, H. S. (2020). Exploring the initial impact of COVID-19 sentiment on US stock market using big data. *Sustainability*, *12*(16), 6648.

Malgarini, M., & Margani, P. (2007). Psychology, consumer sentiment and household expenditures. *Applied Economics*, *39*(13), 1719-1729.

Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change*, *130*, 114-122.

Simionescu, M., & Cifuentes-Faura, J. (2022). Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal. *Journal of Policy Modeling*, *44*(1), 1-21.

Simionescu, M., & Raišienė, A. G. (2021). A bridge between sentiment indicators: What does Google Trends tell us about COVID-19 pandemic and employment expectations in the EU new member states?. *Technological Forecasting and Social Change*, *173*, 121170.

Woo, J., & Owen, A. L. (2019). Forecasting private consumption with Google Trends data. Journal of Forecasting, 38(2), 81-91.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16206

The Role of Twitter and Google Trends in identifying the perception of Russia-Ukraine wars

Vincenzo Miracula, Elvira Celardi
University of Catania, Italy.

Abstract

The COVID-19 pandemic has not only changed the social reality we were used to but also confirmed how data is one of the most valuable resources. We examine the search volume of Google Trends to understand the perception of the war in Ukraine based on people's online information search behaviour and Twitter to figure out how people discuss, react and respond to emergent phenomena from complex events like a war. The data collected from Twitter shows that the public reaction to the events of the 2022 Russia-Ukraine war was diverse, with a large proportion of tweets expressing negative sentiment (≃81%) towards the events. We also show that the use of hashtags such as #NuclearThreat and #RussiaUkraineWar was prevalent during the escalation of the conflict in 2022, indicating that these events were widely discussed on Twitter. The use of these keywords and hashtags can provide a better understanding of how the war is being portrayed in the media and perceived by the general public in pseudo real-time. In order to effectively utilise these data sources, researchers should utilise a combination of quantitative and qualitative methods, including natural language processing and sentiment analysis.

Keywords: *sentiment analysis; google trends; russia-ukraine war; computational social science; behavioural big data*

1. Introduction

In today's globalized and "hyper-digitised" social reality, individuals and groups are constantly leaving trails of their behaviours in real-time, whose extraction and analysis can help to understand the workings of complex social systems and phenomena that have global relevance (such as the impact of recent Ukraine war). The rise of social media, instant messaging, and other forms of digital communication have made it possible for people to interact with each other across vast distances. People are no longer merely consumers of information but become prosumers of content shared in real-time globally. The emerging literature on digitalization (Coleman and Blumler, 2009; Svensson, 2014; Parycek et al., 2017; Tufekci, 2017) highlighted the creation of a "digital agora", a new electronic public sphere that can be seen as a symbol of a more efficient and more emotionally rewarding way to connect citizens and stakeholders (Kamps, 2000, p.228). We think that the use of new technologies in social research might allow us to capture the complexity of the generally unexplored constellations of circumstances that characterise digital contexts. This is a necessary step in intercepting the nonlinear cause-and-effect mechanisms that can result from taking part in debates within the digital agorà. How reflections based on the existing literature and recent international datasets show (Chen and Zhang, 2014; Boyd and Crawford, 2012; Jeble et al., 2017), in fact, by using interconnected data platforms vast amounts of data can be collected and analysed, which can be used to reveal hidden patterns and trends of great utility in numerous decision-making contexts. On the other hand, the use of Big Data (BD) may carry risks related, for example, to their dynamicity, heterogeneity, veracity...validity of information, and, not least, the biases that characterise them. This work grapples with the challenges and opportunities of such technologies for analysing complex social phenomena. A case study related to the current conflict in Ukraine will be presented. More specifically, we examine the search volume of Google Trends to understand the perception of the war in Ukraine based on people's online information search behaviour and Twitter to figure out how people discuss, react and respond to emergent phenomena from complex events like a war. This interaction generates "behavioural data" that captures users' habits in a disaggregated manner (Rhodes et al., 2003; Girardin et al., 2008). In this context, this work also aims to highlight the value such technologies can add to the analysis of complex social phenomena: what features allow these tools to collect and analyse data? What kinds of information do they allow us to capture that more traditional tools cannot?

2. Method

Google Trends and Twitter are useful tools for conducting research. Google Trends allows users to see the relative popularity of search terms over time, providing insight into the public's interest in a particular topic. Twitter, on the other hand, can be used to gather real-time information about a particular event or topic. By searching for keywords or hashtags,

researchers can quickly identify the most relevant tweets and get a sense of how people are discussing a particular issue.

2.1. Google Trends Data

Google search volume is used in several research areas where it is essential to have information about individuals' concerns, interests, and perspectives. In medicine, for example, examining search terms related to flu symptoms has been shown to predict flu activity (Ginsberg et al., 2009). In economics, search volume can be used to predict economic indicators (Choi & Varian, 2012; Da et al., 2011). Finally, during the Covid-19 pandemic, several studies analysed the pandemic situation using search volume (Pan et al., 2020; Walker et al., 2020).

Thus, as has been shown in several research fields, the analysis of search volume can reveal insights into individuals' search for information. Google Trends provides a time series index of the volume of queries users enter into Google. The maximum share of queries in a given time period is normalised to 100. Queries such as "nuclear threat" are counted in the calculation of the query index for "nuclear". Note that the Google Trends data are calculated using a sampling method, so the results vary by a few percentage points from day to day.

We searched "russia-ukraine war" as the first query and we looked at the associated query that had attracted the most interest and found that the words were: sanctions, Nato, Russia, Ukraine, nuclear threat. The data covers a period from February 2022 to December 2022 in order to track the evolution of public interest in the conflict. The data was analysed to identify patterns in the popularity of search terms, as well as any significant spikes or changes in interest over time.

2.2. Twitter Data

Based on what we observed in Google Trends, we decided to study how information about a given event spreads across social networks. In order to get the data for this analysis we decided to collect a significant amount of textual data over time and from a specific social network, Twitter. The reasons behind this choice are fairly easy to understand. Twitter is a social network from which you can easily get data, as it provides a regular API, which is a kind of API (Application Program Interface) that is designed to exchange data over the Internet. In order to get tweets we used Tweepy, a library that can be used through Python code. In order to get data for this work, tweets were collected using a combination of keywords and hashtags related to the 2022 Russia-Ukraine war, such as "Russia-Ukraine" and "Nuclear Threat".

The tweets collected were analysed to identify patterns in the public's reaction to the events, as well as to gain insight into the public's perception of the conflict. Sentiment Analysis was used to classify tweets as positive and negative and Emotion Detection to recognize human

emotion (Ekman, 1992). We ended up collecting 1056313 tweets from European countries filtered by location, language, and hashtag. The data collected from Twitter shows that the public reaction to the events of the 2022 Russia-Ukraine war was diverse, with a large proportion of tweets expressing negative sentiment ($\simeq$81%) towards the events. Many tweets expressed concern and condemnation of the actions of the Russian government, while others expressed support for Ukraine's territorial integrity.

Tweets also showed a high level of condemnation of the use of military force and calls for peaceful resolution. The data also shows that the use of hashtags such as #NuclearThreat and #RussiaUkraine was prevalent during the escalation of the war in 2022, indicating that these events were widely discussed on Twitter. The use of hashtags also allowed users to quickly and easily access information and updates on the conflict. However, it's important to keep in mind that not all Twitter users are credible sources and the information obtained from these sources should be fact-checked and cross-referenced with other sources.

3. Results

The data collected from Google Trends shows that the popularity of search terms related to the Russia-Ukraine war has fluctuated over time (Fig. 1). The data also shows that the popularity of search terms related to the conflict is not limited to Russia and Ukraine, but is also high in other countries such as Germany, France, and Italy. This indicates that the conflict has international resonance and is not limited to the region.

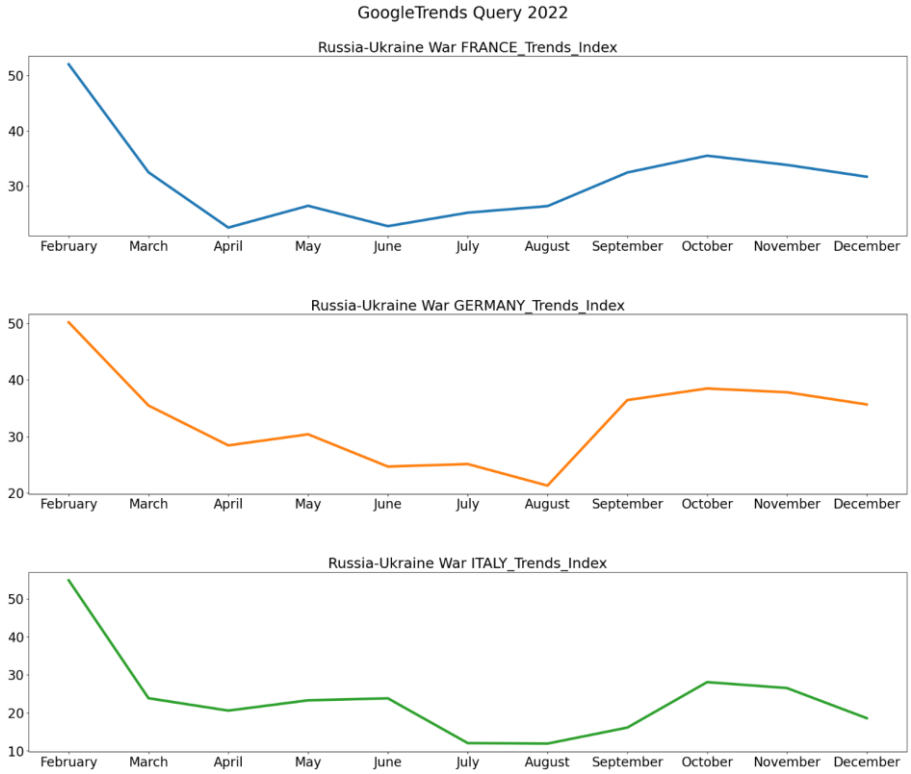


Figure 1. Google Trends Index for countries

As mentioned above, we focused our work on France (Tab.1), Germany (Tab.2) and Italy (Tab.3). For these countries we collected 206.039, 210.976 and 190.946 tweets respectively. Once we obtained the tweets, we conducted a sentiment analysis in order to understand the public opinion. We used RoBERTa (Liu et al., 2019), a pretrained model built on BERT, in order to perform better in the two tasks specific to our objective. In this way, we obtained a sentiment analysis model capable of discerning sentences into two categories: positive and negative. For each of the three countries, we obtained a "Trends_Index" (average of values on google trends), a "Negative_Index" and a "Positive_Index". Each value is expressed as a percentage.

Table 1. France

Months_2022	Trends_Index	Negative_Index	Positive_Index
February	52,07	83,05	16,95
March	32,47	84,20	15,80
April	22,43	84,76	15,24
May	26,39	84,41	15,59
June	22,69	83,64	16,36
July	25,14	84,41	15,59
August	26,33	82,63	17,37
September	32,43	82,78	17,22
October	35,48	83,68	16,32
November	33,81	83,60	16,40
December	31,66	84,33	15,67

Although there is a decline in interest in the conflict in France, high (stable) levels of negative sentiment are observed.

Table 2. Germany

Months_2022	Trends_Index	Negative_Index	Positive_Index
February	52,07	82,13	17,87
March	32,47	84,27	15,73
April	22,43	63,23	36,77
May	26,39	65,62	34,38
June	22,69	83,07	16,93
July	25,14	83,24	16,76
August	26,33	83,50	16,50
September	32,43	83,81	16,19
October	35,48	83,39	16,61
November	33,81	83,94	16,06
December	31,66	83,69	16,31

There is also a decline in interest in the conflict in Germany, but different levels (April and May) of negative sentiment are observed.

Table 3. Italy

Months_2022	Trends_Index	Negative_Index	Positive_Index
February	54,85	83,85	16,15
March	23,85	83,44	16,56
April	20,59	83,82	16,18
May	23,30	82,96	17,04
June	23,82	83,35	16,65
July	12,05	63,77	36,23
August	11,92	52,97	47,03
September	16,13	53,02	46,98
October	28,08	83,08	16,92
November	26,52	83,71	16,29
December	18,60	53,39	46,61

In Italy there is a sharp decline in interest in the war, especially in the summer months and in December. In these moments the sentiment values are almost equal.

4. Conclusion

The use of Google Trends and Twitter data can provide valuable insights into the public perception and discourse surrounding the Russia-Ukraine war in 2022. By analysing the frequency and sentiment of certain keywords and hashtags related to the conflict, researchers can gain a better understanding of how the war is being portrayed in the media and perceived by the general public in pseudo real-time. Additionally, tracking the geographic location of tweets and search queries can provide insight into which regions and countries are particularly engaged with the war. It is also important to consider potential biases and limitations of the data, such as the fact that social media usage and access to the internet may not be representative of the entire population. The tools presented in this paper provide useful support in reconstructing the contextual conditions in which complex social phenomena develop, like the one examined in the study presented here.

These tools are useful for social research because they allow for the reconstruction - through the collection of large amounts of data - of the digital and globalised contexts in which

information spreads. That way, trends can be observed and reactions and behaviours can be hypothesised, with a speed that traditional research tools cannot. To understand what mechanisms are triggered by certain contextual conditions, leading individuals to put in place specific behavioural responses, it is considered necessary, however, to supplement the information found online with others related to the historical, cultural, social, technological…dimensions of the real contexts. To obtain other relevant findings besides the one presented through this exploratory phase of the research, the information collected will be supplemented with additional quantitative data. It would also be necessary to include in the analysis the particular point of view of the social actors who populate the real contexts. This element escapes the analysis of big data.

References

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, *275*, 314-347.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic record*, *88*, 2-9.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The journal of finance*, *66*(5), 1461-1499.

Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-1014.

Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. Pervasive Computing, IEEE, 7(4), 7885.

Jeble, S., Kumari, S., & Patil, Y. (2017). Role of big data in decision making. *Operations and Supply Chain Management: An International Journal*, *11*(1), 36-44.

Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological forecasting and social change, 130, 69-87.

Mejias, U. A., & Vokuev, N. E. (2017). Disinformation and the media: the case of Russia and Ukraine. Media, culture & society, 39(7), 1027-1042.

Pan, Z., Nguyen, H. L., Abu-Gellban, H., & Zhang, Y. (2020, December). Google trends analysis of covid-19 pandemic. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 3438-3446). IEEE.

Rhodes, S. D., Bowie, D. A., & Hergenrather, K. C. (2003). Collecting behavioural data using the world wide web: considerations for researchers. Journal of Epidemiology & Community Health, 57(1), 68-73.

Walker, A., Hopkins, C., & Surda, P. (2020, July). Use of Google Trends to investigate loss of smell–related searches during the COVID-19 outbreak. In *International forum of allergy & rhinology* (Vol. 10, No. 7, pp. 839-847).

Zhang, Z., Zhang, Y., Shen, D., & Zhang, W. (2018). The cross-correlations between online sentiment proxies: Evidence from Google Trends and Twitter. Physica A: Statistical Mechanics and Its Applications, 508, 67-75.

Two stories of cancel culture. How value-driven calls to cancel affect the bottom line

Paul Reyes-Fournier[1], Elizabeth Reyes-Fournier[2], David Bracken[3]
[1]Psychology Department, Keiser University, USA, [2]Psychology Department, Keiser University, USA, [3]Psychology Department, Keiser University, USA.

Abstract

Social media has become the environment for value-driven polemics that have been commonly dubbed cancel culture. As a construct, cancel culture has little empirical research especially as it relates to the efficacy of the calls to cancel a business. This research analyzes a successful call to cancel, evidenced in a break in sales for Abercrombie & Fitch, and an unsuccessful call to cancel directed toward Starbucks. The sentiments of tweets for both companies were reduced to a valenced level for the core emotions, using a one-dimensional k-mean clustering, for each fiscal quarter. Correlational and time-series analysis was performed. A successful call to cancel showed a structural break in sales but not in Altman's z-score. The polemic differences in the emotions were strongly correlated to sales for the successful call to cancel but were not present in the unsuccessful case. Likewise, the time-series analysis showed Granger-causality between emotions and the sales for a successful call to cancel. In both the successful and unsuccessful campaigns, individual emotions were ultimately found to be representing two factors present in the 8-emotion model- positive and negative sentiment. This research supports the threshold model of consumer decision-making while calling into question the granular nature of emotions.

Keywords: *Cancel Culture; Theory of Emotion; Sentiment Analysis; Time Series Analysis; decision-making in purchasing*

Google Search Volume Index: A systematic Review

María José Ayala[1], Nicolás Gonzálvez-Gallego[1], Rocío Arteaga-Sánchez[2]
[1]Department of Business Administration, Catholic University of Murcia, Spain,[2]Department of Business Administration and Marketing, University of Seville, Spain.

Abstract

Information is a crucial and key element when studying stock markets and the way it is analyzed can be determinant for measuring financial market movements. With Internet uptake, investors are exposed to a vast amount of information and hence, analyzing what they search can provide relevant data about potential investment actions and trading decisions. In other words, measuring what investors search yields information about how present and future assets prices change. Recently, the research community has focus on measuring investor attention through search queries on Google. In this manner, investor attention is considered as the frequency of a specific term searched, presented by Google Search Volume Index (GSVI).

This paper conducts a systematic review of the current literature about the use of GSVI as a proxy variable for investor attention and stock market movements explanations. Using Web of Sciences and Science Direct data bases, we analyze 51 academic studies published between 2010 and 2021. The articles are classified and synthetized based on the selection criteria for building GSVI: keyword of the search term, market region and frequency of the data sample. After that, we analyze the effect over the financial variable Return, Volatility and Trading volume for measuring the effect of GSVI over market movements. The main results can be summarized as follows: (1) GSVI is positively related with volatility and trading volume regardless the keyword, market region or frequency used for the sample. Hence, an increase on investor attention toward a specific financial term will lead to an increment on volatility and trading volume; (2) GSVI can improve forecasting models for stock market movements. To conclude, this paper consolidates for the first time the research literature about GSVI, being highly valuable for academic practitioners of the area.

Keywords: *Google Trends, Investor attention, GSVI, stock market prediction.*

Analysing the process of territorial data collection for the Consumer Price Survey

Loredana De Gaetano[1], Gabriella Fazzi[2], Serena Liani[3]
[1] Data Collection Directorate, Italian Institute of Statistics (Istat), Italy [2] Data Collection Directorate, Italian Institute of Statistics (Istat), Italy [3] Data Collection Directorate, Italian Institute of Statistics (Istat), Italy.

Abstract

In order to improve the process of territorial data collection for the Consumer Price Survey, the Data Collection Directorate of the Italian Institute of Statistics (Istat), in collaboration with the Istat experts, has undertaken a survey of the Municipality Statistics Managers who did not take part in the process. Listening to the point of view of the stakeholders who are directly involved in the data collection process is a necessary starting point for a analysis of the design of a territorial survey. In fact, it prevents solutions being imposed from above, which could be ineffective, far from the real needs of those who are asked to collect reliable and timely data. The information gathered allows the planning of changes in terms of a modernization of the data collection mode and a leveraging of new data sources (administrative and big data).

Keywords*: data collection; mixed-mode; new data sources.*

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16467

Density modelling with functional data analysis

Stefano A. Gattone[1], Tonio Di Battista[1]
[1]DISFIPEQ Department, University "G. d'Annunzio" of Chieti-Pescara, Italy

Abstract

Recent technological advances have eased the collection of big amounts of data in many research fields. In this scenario density estimation may represent an important source of information. One dimensional density functions represent a special case of functional data subject to the constraints to be non-negative and with a constant integral equal to one. Because of these constraints, a naive application of functional data analysis (FDA) methods may lead to non-valid results. To solve this problem, by means of an appropriate transformation densities are embedded in the Hilbert space of square integrable functions where standard FDA methodologies can be applied.

Keywords: *Bayes space; Density; Functional data analysis; Transformation approach.*

1. Introduction

This work deals with density modeling using functional data analysis. (Ramsay and Silvermann, 2005). In scenario with big amounts of data collection, probability density functions can provide more information than single summary statistics.

One of the main goals in statistical data analysis is to associate the change of (a function of) some response variable *y* with a set of covariates *x*. The most common tool for this is mean regression which focuses on the conditional expectation *of y given x*. Quantile regression models investigate specific quantiles of the conditional distribution of the response. By modelling the entire probability distribution of the response, density regression methods consider the impact of the covariates on the entire distribution. Densities could represent the data atoms of interest such as yearly income distribution, population age and mortality distributions across different countries.

Probability density functions (pdfs) represent a special case of functional data since they must satisfy the constraints of being non-negative everywhere and present a constant integral equal to one. Standard functional data analysis (FDA) methods cannot be naively applied without considering such constraints. To address this issue several strategies can be found in the literature.

One strand of literature represents densities as elements of the so-called Bayes space starting from the Aitchison geometry valid for compositional data (Aitchison, 1982). In this setting, pdfs are represented by a centred log-ratio transformation which represents an isometric isomorphism between the Bayes space of pdfs and the Hilbert space (Hron et al, 2016).

Another approach is envisaged by Petersen and Muller (2016) where the pdfs are mapped into a linear functional space through a suitably chosen transformation. Established methods for Hilbert space valued data can be applied to the transformed functions and the results are moved back into the density space by means of the inverse transformation. Examples of transformations are the log-hazard transformation and the log-quantile density transformation. The view is completed by considering the objected-oriented analysis of densities where spaces are equipped with metrics such as the Wasserstein or the Fisher-Rao providing a manifold structure on probability distributions. Within this framework, tangent space structures need to be defined to facilitate computations (Petersen and Muller, 2019).

2. Density functions as constrained functional data

A functional variable is defined as a random variable *f*, taking values in an infinite functional space, the Hilbert space of square integrable functions equipped with the usual inner product and norm:

$$H(t) = \left\{ f: T \to \mathbb{R} \ such\ that \int_T f(t)^2\, dt < \infty \right\} \quad (1)$$

with $\langle f, g \rangle = \int f(t)g(t)\, dt$ and $\|f\| = \left\{\int_T f^2(t)\, dt\right\}^{\frac{1}{2}}$.

We are interested in the case where the observed functions are density functions. We denote with *D* the functional space of density functions. In this space functions are positive and integrate up to 1 as described in equation (2):

$$D(t) = \left\{ f: T \to \mathbb{R} \ such\ that\ f(t) > 0 \text{ and } \int_T f(t)\, dt = 1 \right\} \quad (2)$$

We assume the data consists of a sample of *n* random density functions. In many situations, the densities themselves will not be directly observed. Instead, a sample of data that are generated by the random density is available. Thus, there are two random mechanisms at work: the first generates the sample of densities and the second generates the samples of data. Typically the first step in working with functional data is the use of basis expansion and penalized smoothing. Estimation is developed, for example, in the natural cubic splines framework:

$$\sum_j \left[y_j - f(t)\right]^2 + \lambda \int [D^2 f(t)]^2 dt \quad (3)$$

where y_j are the observed discrete data points which must be converted to a functional data object *f*. The constant lambda is the smoothing parameter with larger values resulting in smoother fits. Now, imagine imposing on the estimated function f some constraints. The constrained curves cannot be treated as vectors in the Hilbert space since a plain basis expansion of the curves does not guarantee the fulfilment of the constraints. In other words, the problem is to simultaneously smooth nonlinear structure in data and incorporate constraints.

3. The w-transform

Let Y have an un unknown positive density function. Following Ramsay and Silvermann (2005) we can write its log-density function in the form $w - C(w)$ where

$$C(h) = log \int exp[w(y)] dy \quad (4)$$

The corresponding log-likelihood function is given by

$$l(w, \boldsymbol{Y}) = w(\boldsymbol{Y}) - C(w) \tag{5}$$

Note that $w(\boldsymbol{y})$ is not constrained in any way. In this way a constrained problem is transformed into an unconstrained one that reduces to the modelling of $w(\boldsymbol{y})$. The modeling of $w(\boldsymbol{y})$ can be obtained by using a flexible nonparametric estimator based on spline basis functions. Once the estimator is obtained, we are able to map the densities into the Hilbert space since the functions $w(\boldsymbol{y})$ are free of constraints. Our proposal is to apply linear FDA methods in the transformed linear space and eventually results on the linear space are mapped back into the density space by means of an appropriate inverse map.

4. Applications

In many application fields, densities are the data atoms of interest such as yearly income distribution, population age and mortality distributions across different countries or distribution of cross-sectional financial returns of different firms or different markets.

Data analysis frequently concerns itself with associating the change in a function of some response variable *y* with a set of covariates *x*. The most common tool for this is mean regression which focuses on the conditional expectation *of y given x*. This prevents inference about other parts of the conditional density. Quantile regression models investigate specific quantiles of the conditional distribution of the response. In such circumstances, individual quantiles are being targeted as proxies of the distribution. By modelling the entire probability distribution of the response, density regression methods perform a substantially harder task than mean and quantile regression. In doing so, one can consider the impact of the covariates on the entire distribution.

A naïve application of the function-on-scalar regression or the function-on-function regression model (Ramsay and Silvermann, 2005) would not guarantee the estimated response to fulfill the definition of a density. Similarly, to compositional regression (Talskà et al., 2018), an alternative could be applying the functional regression model on the unconstrained functions w(t) in eq. (5) and then the parameters estimates are mapped back to the density space applying the inverse transformation. In contrast to the estimates resulting from the naïve functional regression model, the estimates are bona fide density function.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44, 139-177.

Hron, K., Menafoglio, M. Templ, M., Hruzova, K. & Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes space. *Computational Statistics & Data Analysis*, 94, 330-350.

Petersen, A., & Muller, H. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 32 (1), 183-218.

Petersen, A., & Muller, H. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47 (2), 691-719.

Ramsay, J.O. & Silverman, B.W. (2005). *Functional Data Analysis*, 2nd edn. New York: Springer.

Talskà, R, Menafoglio, A., Machalovà, J., Hron, K. & Fiserovà, E. (2018). Compositional regression with functional response. *Computational Statistics and Data Analysis*, 123, 66-85.

Assessing the impact of innovation signaling on the investment

Mikaël Héroux-Vaillancourt[1], Catherine Beaudry[1], Davide Pulizzotto[1], Margaret Dalziel[2]

[1]Canada Research Chair in Management and Economics of Innovation, Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Canada,[2] Conrad School of Entrepreneurship and Business, University of Waterloo, Canada.

Abstract

This exploratory study investigate the use of innovation-related language in corporate website from a signaling perspective. We empirically tested whether the occurrences of innovation-related terms in corporate websites contributes to the investment received by the firms. With a sample of 1,289 firms who participated in 21 questionnaire-based investigations between 2010 to 2016, we extracted the content of the corresponding websites via snapshots hosted on The Wayback Machine. We built indicators based on a document frequency analysis of the keywords related to various innovation factors (innovation culture, collaboration, open innovation, R&D and IP). The OLS regression shows that innovation-related signaling on corporate websites is significantly related with the investment received. The study contributes in understanding the intention behind the use of innovation-related signaling in the corporate world and proposes a new indicator to identify innovation-active firms that are seeking external support.

Keywords: *innovation; investment; market signaling; web mining; Document keyword frequency analysis; OLS regression.*

Preventing data quality issues with data contracts: a proactive solution

Vanya Petrova Kostova
Global TECH, HelloFresh SE, Germany.

Abstract

Data quality is a critical aspect of data product management and a major challenge in the field of data engineering. It refers to the availability, accuracy, completeness, and consistency of data, which are essential factors for reliable and informed data-driven business decisions. In data flows, data quality is often compromised by errors, missing values and inconsistencies that occur already in the initial source systems. In practice, such issues are getting addressed with filtering and post processing techniques to clean and format the data accordingly but delays from the time of detection until resolution may cause unacceptable risks in data-driven decision-making processes and thus, may harm the overall business.

We propose data contracts as a mechanism to address data quality issues at the root cause. Data contracts between data producers and data consumers enable defining and tracking data lineage and ensure that a data consumer can analyse and model the data in time-critical business situations. Compared to the more traditional approaches of data quality monitoring and alerting, which are designed to identify and raise an issue, data contracts can help organisation to avoid data quality issues before they affect data flows and business operations.

Keywords: *data quality; data contracts; data products; data-driven decision-making.*

Editorial Universitat Politècnica de València

Evaluation of term-weighting measures for grouped text documents with a target variable: a simulation study

Riccardo Ricciardi[1], Marica Manisera[1]
[1]Department of Economics and Management, University of Brescia, Italy.

Abstract

In Text Mining applications, count-based models are often used to represent text documents. When two document variables are available, i.e. an outcome and a grouping variable, the weight of a word for the documents may depend on the group memberships. The contribution of this work is to frame this context with a statistical approach, by modelling the corpus of documents with a Multivariate Binomial distribution (Hudson et al., 1986). The advantage of this solution is two-fold: it allows (1) to review, in a statistical framework, some term-weighting measures used in the literature (Samant et al., 2019), and (2) to simulate corpora with predefined characteristics by means of the Gaussian Copula method (Genest and McKay, 1986). This simulation is useful to investigate the ability of the existing measures, computed on the group-word interaction, to capture both the group-word relationship itself and the target-word association. Results from the simulation study show interesting relationships that can be exploited by nice visualization tools.

Keywords: *Term-weighting measures; Gaussian Copula; Simulation.*

References

Hudson, William N., Howard G. Tucker, e Jerry A. Veeh. 1986. Limit Theorems for the Multivariate Binomial Distribution. *Journal of Multivariate Analysis* 18(1):32–45.

Genest, Christian, e Jock MacKay. 1986. The Joy of Copulas: Bivariate Distributions with Uniform Marginals. *The American Statistician* 40(4):280–83.

Samant, Surender Singh, N. L. Bhanu Murthy, e Aruna Malapati. 2019. Improving Term Weighting Schemes for Short Text Classification in Vector Space Model. *IEEE Access* 7:166578–92.

Discourses as units of knowledge in the light of neural language models. Refinement of the theory of discursive space

Rafal Maciag[1]
[1]Institute of Information Studies, Jagiellonian University, Poland

Abstract

In recent years, and even months, a rapid development of NLP solutions can be observed. This technology allows one to define deeper semantic inferences in the text based on the idea of neural language models (NLMs). Neural language models (NLMs) are containers of knowledge. The relationship between language and knowledge has been extensively reflected in research in the form of the so-called discourse analysis. Based on Michel Foucault's concept of discourse, especially the text from 1971(Foucault, 1971), knowledge model was proposed named discursive space, in which discourses as instances of knowledge travel trajectories in a multidimensional dynamical space (Maciag, 2022). The idea presented in the paper assumed that it is possible to isolate semantic structures more complex than the semantic units used so far, i.e. tokens, which are based on words and their relationships in sentences. Such structures are discourses, i.e. linguistic (semantic) structures with a higher degree of abstraction than the sentences they consist of. Therefore, one should search for higher-order units (discourses) composed of lower-order semantic units (words) and their relations in sentences. It would be a repetition of the embedding technique used in NLM, but transferred to a higher semantic level, the aim of which is to create a set of vectors describing discourses. By analyzing the mutual position of the indicated discourses in the corpus of texts, a discursive linguistic model would be created. The introduction of a variable in the form of time, i.e. the construction of a dynamical discursive model, would fulfill the assumptions of discursive space.

Keywords: *natural language processing; neural language models; knowledge; discourse; discursive space.*

Foucault, M. (1971). *L'ordre du discours: Leçon inaugurale au Collège de France prononcée le 2 décembre 1970*. Paris: Gallimard.

Maciag, R. (2022). Theory of Knowledge Based on the Idea of the Discursive Space. *Philosophies*, *7*(4), 72. doi: 10.3390/philosophies7040072

Suitability of various machine learning approaches for recognition of antisocial behaviour on social networks

Kristína Machová[1], Tomáš Tomčík[1]
[1]Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Slovakia.

Abstract

Nowadays, social networks allow web users to express publicly agreement or disagreement with other people and express freely their opinions. This freedom is often abused and that is why we can see social networks that are full of offensive comments. The increase in textual data on the Internet has stimulated the emergence of new scientific fields as web mining that examine short texts in the online space and look for hate or offensive speech, and that try to analyze textual data in online space. Our paper is focused on a special type of analysis concentrated on detection of some forms of antisocial behaviour, particularly on hate speech, offensive posts, and cyberbullying recognition in the online space. The main goal of the work was to find out which of the machine learning strategies - classic, deep or ensemble - are the most effective in detecting of these forms of antisocial behaviour on social networks. We have compared models generated by the following methods: deep learning of neural networks (LSTM, and GRU), classical methods (SVM, NB, and DT), and ensemble learning (RF, AdaBoost). We have tested those methods on three datasets created from posts of various volume to find how the volume of data available for training affects the results of machine learning models. The best result on the smallest Hate Speech Dataset were achieved by ensemble learning using AdaBoost (Accuracy=0,904). On the other hand, the best result on the largest Offensive Speech Dataset was achieved by deep learning using GRU (Accuracy=0.964).

Keywords: *Machine learning; deep learning; ensemble learning; social web mining; detection of antisocial behaviour.*

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16457

Mapping policymaker narratives of the Climate Security nexus on Social Media: a case study from Kenya

Bia Silveira Carneiro[1], Giuliano Resce[2], Giosuè Ruscica[1], Giulia Tucci[1]
[1]Alliance of Bioversity International and CIAT, Portugal [2]Department of Economics, University of Molise, Italy

Abstract

Despite increasing awareness of the nexus between climate change and human security, especially in fragile contexts, this complex relationship has yet to be reflected in the policy arena. To investigate this potential policy gap, we apply an online issue mapping approach to assess representations of climate security within the public discourse of policymakers on social media, using Kenya as a case study. Considering Twitter as a proxy for public debate, text mining and network analysis techniques were employed to a corpus of almost 50 thousand tweets from selected national-level state actors, aiming to identify the evolution of thematic trends and actor dynamics. Results show a disassociation between climate, socioeconomic insecurities, and conflict in the public communications of national policymakers. These findings can have useful implications for the policy cycle, indicating where policy attention around climate security-related topics has been and what are the entry points for enhancing sensitivity on the issue.

Keywords: *climate security; social media; text mining; sentiment analysis; online issue mapping; digital methods.*

1. Introduction

Climate security refers to climate-related threats to societies, communities and individuals that encompass risks directly or indirectly caused by climate change, including the potential for conflict. The relationship between climate and conflict has been receiving increased attention in the past decade, as the climate crisis has been shown to impact social and political stability. However, despite heightened awareness regarding potential linkages between climate, peace and security, such connections have yet to be reflected in the policy arena. Policy cycles for climate adaptation and mitigation, as well as national security concerns, often fail to reflect the complex pathways that link the two dimensions (Brzoska 2012).

To explore the potential policy gap related to the climate-security nexus, a data-driven method was developed to assess representations of climate security as a topic of governance within the public discourse of state actors, using Kenya as a case study. While the country is characterised as relatively peaceful compared to some of its neighbours, climate variability and extremes have had adverse impacts on agricultural production. Compounded by external shocks that exacerbate existing inequalities, Kenya faces increased risks of resource-related violence (CGIAR FOCUS Climate Security, 2022). Consequently, though climate may not directly impact localized conflict dynamics, its context-specific interactions with socio-economic and political factors can shape and increase risks of human insecurity and conflict.

While extensive research about climate change discourses on social media have been conducted, focusing on various subjects such as issue polarization, disinformation, activism, and climate communication (see Pearce et al 2019 and Falkenberg et al 2022, among many others), this study investigates this study relies on the foundations of Digital Methods (Rogers 2013; Carneiro et al 2022) to explore climate security narratives and dynamics among policy actors. An online issue mapping approach (Rogers et al. 2015) was applied to investigate two main questions: 1) How salient are climate security issues in national policy agendas? 02) How are linkages between climate, socioeconomic risks and insecurities, and conflict represented in the public narratives of policymakers? Insights emerging from this analysis provide a starting point for the development of evidence-based advocacy and engagement strategies so that effective responses to climate change are sensitive to the interlinkages with the human security context in the country.

2. Methods and data

Twitter has been widely recognized as an important venue for institutional communications; news media increasingly rely on the platform as a primary source of official statements and positiontaking. Its potential as a real-time, topic-driven platform enables rapid detection of trends to uncover discourse dynamics (McDonald 2013). Hence, to frame perceptions around the climate-socioeconomic insecurities-conflict nexus at the national policy level in Kenya,

an analysis of government communications on Twitter was performed. An algorithm was developed to extract all publicly available Tweets from the official accounts of central government bodies, ministries of agriculture, environment, and natural resources, as well as national security bodies (Table 1), from which the presence of a climate security taxonomy was explored. In total, 49,335 Tweets were collected between 2012-09-13 to 2022-05-26.

Table 1: Categories of official Twitter accounts of state actors selected for analysis.

Categories of Twitter Accounts	No. Tweets
Central government	12850
Ministries of agriculture	1040
Ministries of environment	7379
Ministries of natural resources	8863
Security-related bodies	19203
TOTAL	49335

A scoping review on the climate security nexus in Kenya (Dutta Gupta et al., 2021) identified 111 topics organised into a framework with five categories: climate; conflict; livelihood and food security; resource availability and access; state capacity and resource governance. Based on this framework, a custom taxonomy was created using the term expansion strategy proposed in Carneiro et al (2022), in which topics were matched to AGROVOC[1], the Food and Agriculture Organization's (FAO) open-source, multilingual vocabulary. For each topic, the corresponding AGROVOC concept was extracted, and a custom algorithm was developed to detect and classify the related terminology within the text of Tweets. Topics were then assessed through correlation measures to identify any interlinkages.

In addition, leveraging on the specific affordance of Twitter that enables direct conversations among users, a network analysis assesses the relationships among policy actors through a mentions network (Williams et al 2015), where accounts are the nodes and their relations are the lines connecting pairs of nodes. This means that accounts are connected if they are mentioned by another, with the weight of the connection calculated from the number of mentions by the same account.

3. Results

Drawing on the mechanisms through which climate stressors may interact with socioeconomic, ecological, and political dimensions identified in Kenya, figure 1 shows their overall distribution, as frequency counts. 'Famine' and 'Aid programmes' are the most regularly mentioned topics, followed by resource availability and access pathway variables

[1] https://www.fao.org/agrovoc/

'Ethnic groups' and 'Cattle'. The most present topics for climate variables are 'Risk' and 'Weather hazards' and for conflict variables are 'Crime' and 'Sexual violence'.

Figure 1 Distribution of topics identified in Tweets from the official accounts of selected government bodies. More frequent terms represented by wider wedges in the pie chart.

While the overall distribution of variables uncovers the cumulative prominence of topics, a temporal distribution provides a more nuanced perception of topic prevalence over time. Beyond the presence or absence of a topic, the algorithm also quantified their presence2. Figure 2 presents timelines for the prevalence of climate variables (top) and conflict variables (bottom) in the corpus of tweets. The visualisation indicates not only which topics were in focus, but also when they were of most interest. Among climate variables, 'Drought' presents several major peaks. In 2018 and 2022, they reflect consecutive failed seasons that led the Kenyan government to declare a national disaster in several parts of the country in 20213. The conflict timeline shows higher variability among topics, with 'Armed conflict', 'Dispute', 'Theft', and 'Crime' oscillating between peaks and low prevalence. As noted by DuttaGupta et al (2022), 'Theft' is most likely associated to livestock raiding, a significant problem in the country's rural areas, whereas the increase in 'Disputes' in the last five years points to increased attention to conflict over resources.

[2] Values were normalized on a scale from 0-1, so that prevalence is shown as a proportion of all frequency, on all topics, in the corpus of Tweets.

[3] https://www.businessdailyafrica.com/bd/economy/kenya-declares-drought-national-disaster-3543276

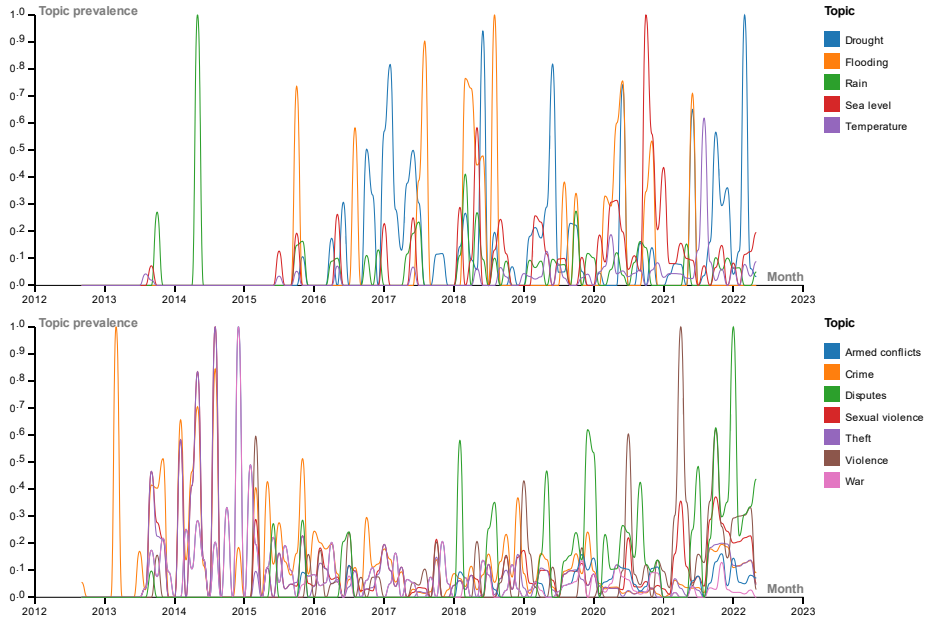


Figure 2 Timeline of Tweets that contain climate (top) and conflict (bottom) variables.

To further unpack the interlinkages between different topics within the Tweets, a measure of correlation was established to identify when terms are present within the same body of text. A strong positive correlation indicates that the terms consistently occur within the same Tweet, whereas a negative correlation denotes they are occurring in separate Tweets. Figure 3 displays the 10 topics most positively correlated to climate variables (right) and to conflict variables (left). In the case of climate, the strongest associations are to livelihood and food security pathway and resource availability and access pathway variables; conflict is not represented in the table. Conversely, conflict-related content is frequently co-occuring with several climate and socioecological variables, namely 'Desertification', 'Risk', 'Climate change, 'Poaching', 'Environmental degradation', and 'Resource management'.

Direct associations between climate and conflict are presented in Figure 4, which features the correlations among the six conflict types described in the impact pathways and climate stressors and socioecological variables. The strongest positive associations (in blue) concern 'Disputes' with 'Erosion', 'Theft' with 'Rain', and 'Violence' with 'Environmental degradation'. However, it is notable that most variables present negative associations (in red), meaning that the connection between climate and conflict is largely absent from the official discourse of Kenyan government actors on Twitter.

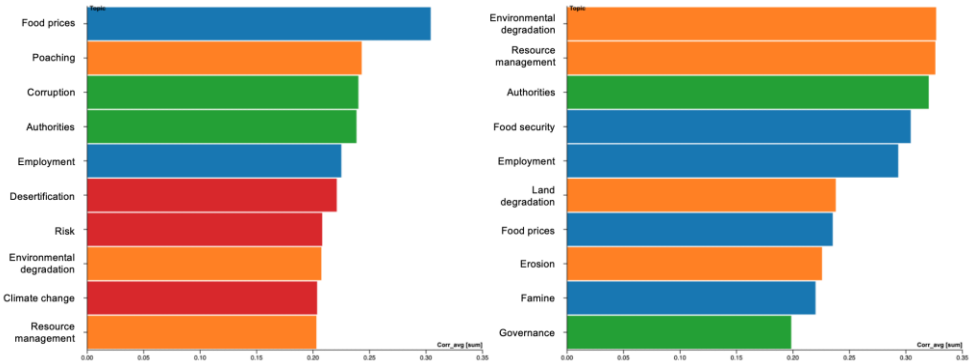


Figure 3 Top 10 correlations between climate (left) and conflict (right) with other topics identified the Tweets extracted from the official accounts of selected government bodies. Bar coloured by category: Climate (red), Livelihood and food security pathway (blue), Resource availability and access pathway (orange), State Capacity and resource governance pathway (green)

Lastly, the mentions network in figure 6 presents a visualization of the dynamics between the key state bodies representing climate, agriculture, natural resources and security interests. As the focus of this analysis is on the interaction between national security and climate adaptation and mitigation policy narratives, the network was filetered to display only the connections between the 13 government accounts. The spatialization of nodes was estimated with the force-directed algorithm Fruchterman-Reingold (Fruchterman and Reingold 1991) , which moves nodes further or closer from each other in an attempt to find an equilibrium. The sizes of the nodes and the labels are partitioned by degree centrality, a measure of the number of connections to a particular node, whereas the edges are also weighed by the number of times a pair of nodes is connected. The graph points to the strongest connections between the ministries and central government accounts, but to weaker or non-reciprocal linkages between ministries from the different areas.

4. Discussion and conclusion

Content analysis and network analysis techniques enable identification of trends in political agendas and actors over time. The machine-driven approach employed to explore the salience of climate security in the Twitter communications of Kenyan policy actors found that the pathways that link climate stressors, socioeconomic risks, and conflict are not well represented in the narratives of government bodies. While Tweets that addressed different types of conflict did show some association to ecological threats, most climate and conflict variables were negatively correlated. Further, the weaker or absent connections in the network analysis point to to potential gaps in dialogue.

A limitation of our analysis is that social media narratives may not fully capture the complexity of policy cycles in a country like Kenya, where policy actors interact across

multiple scales, and this engagement may not be adequately represented in digital platforms. Moreover, the African continent continues to have the lowest Internet access, with Kenya's internet penetration rate at 32.7 percent of the total population at the start of 2023, and Twitter reaching 3.5 percent of the total population4. However, given the continued trend to integrate digital platforms in policy and governance, especially during times of crisis such as the Covid-19 pandemic or natural disaster responses, this study contributes towards mapping policymaker perspectives in public discourse. Our findings can have useful policy implications, indicating where policy attention around climate security-related topics has been, as well as what are the gaps and entry points for enhancing sensitivity on the issue, facilitating the integration of the climate security debate into Kenya's formal policy arena.

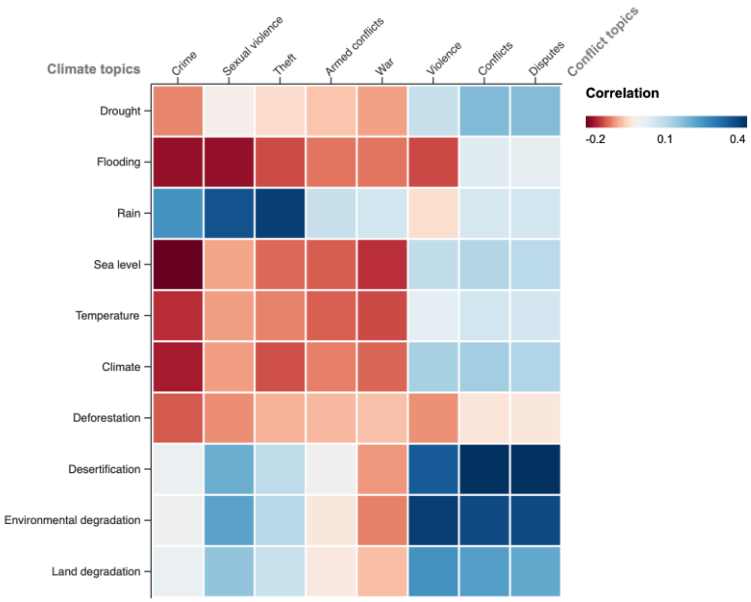


Figure 4 Correlations between conflict types and climate shocks identified in the Tweets extracted from the official accounts of selected government bodies.

[4] https://datareportal.com/reports/digital-2023-kenya (retrieved 19 April 2023)

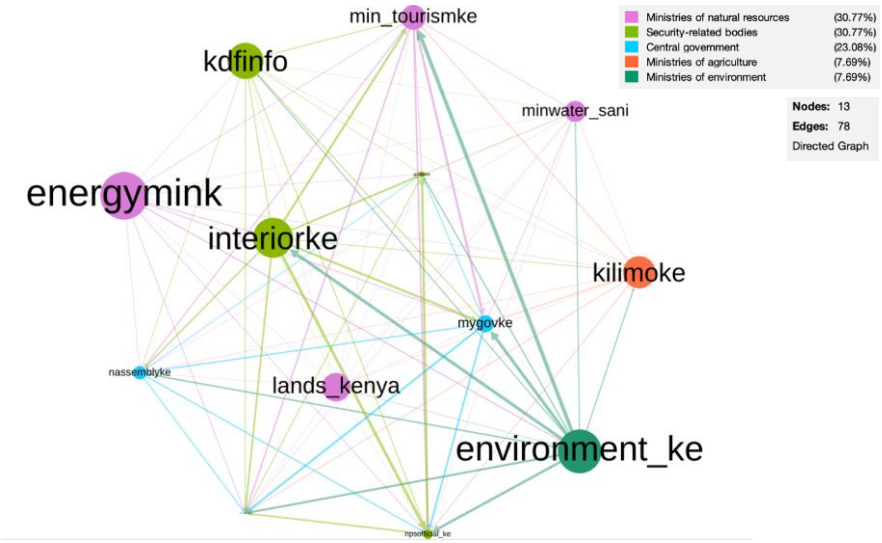


Figure 5 Network of policy actors.

References

Brzoska, M. (2012). Climate Change as a Driver of Security Policy. In: Scheffran, J., Brzoska, M., Brauch, H., Link, P., Schilling, J. (eds) *Climate Change, Human Security and Violent Conflict.*, vol 8. Springer, Berlin, Heidelberg.

Carneiro, B., Resce, G. & Sapkota, T.B. (2022). Digital artifacts reveal development and diffusion of climate research. *Scientific Reports*, 12, 14146.

CGIAR Focus Climate Security (2022) *Climate Security Observatory Country Profile: Kenya. Rome*, Italy: CGIAR FOCUS Climate Security.

Dutta Gupta, T., Madurga Lopez, I., Läderach, P., & Pacillo G. (2021). *How does climate exacerbate root causes of conflict in Kenya? An impact pathway analysis.* Rome, Italy: CGIAR FOCUS Climate Security.

Falkenberg, M., Galeazzi, A., Torricelli, M. *et al.* (2022) Growing polarization around climate change on social media. *Nat. Clim. Chang. 12*, 1114–1121.

Fruchterman, T.M.J. and Reingold, E.M. (1991), Graph drawing by force-directed placement. Softw: Pract. Exper., 21: 1129-1164.

McDonald, M. (2013). Discourses of climate security. *Political Geography*, 33, 42–51.

Pearce, W, Niederer, S, Özkula, SM, Sánchez Querubín, N. The social media life of climate change: Platforms, publics, and future imaginaries. *WIREs Clim Change*. 2019; 10:e569.

Rogers, R. (2013) *Digital Methods*, The MIT Press.

Rogers, R., Sanchez-Querubin, N., & Kil, A. (2015). *Issue Mapping for an Aging Europe*, Amsterdam: Amsterdam University Press.

Williams, H. T. P., McMurray, J. R., Kurz, T. & Hugo Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change* 32, 126–138

Applying transformers-based NLP models to explore credibility in different product categories in Amazon's online reviews

María Olmedilla[1], José Carlos Romero[2], Rocío Martínez-Torres[3], Sergio Toral[4]
[1]SKEMA Business School – Université Côte d'Azur, France, [2]Applied Computational Social Sciences-Institute, University of Paris-Dauphine-PSL, France, [3]Facultad de Ciencias Económicas y Empresariales, University of Seville, Av. de Ramón y Cajal, 1, 41018 Sevilla, [4]E. S. Ingenieros, University of Seville, Avda. Camino de los Descubrimientos s/n, 41092, Seville, Spain.

Abstract

Online reviews in the e-commerce and eWOM communities play a key role in consumers' purchase decisions. In this regard, one concern is the growth of fake reviews, which directly targets the credibility of platforms and the trust of users. To address this issue, we apply Transformers-based NLP models to better understand the scope of fake reviews within the Amazon marketplace across different product categories. Our methodology applies two different transformer models to Amazon online reviews for (1) generating fake reviews and (2) classifying online reviews as fake or truthful. This work contributes to the literature on understanding the credibility of online review. Our results show that most of the fake reviews are located in non-verified purchase reviews. Considering the different product categories, we found that the percentage of fake reviews is 3 times higher for the experience products and 8 times higher for the experience products for non-verified purchase reviews with respect to the fake reviews found in verified-purchase reviews.

Keywords: *online reviews; transformers; GPT-2; BERT; credibility; verified purchase*

This work was supported by the project Aplicación de Redes Generativas Antagónicas para Combatir la Manipulación de Clientes Online (REACT) Ref. PID2020-114527RB-I00 funded by MCIN/AEI/10.13039/501100011033

Engagement analysis on Instagram: contributions to the co-creation of tourism experiences

Marta Andrade-Cunha[1], Ana Irimia-Diéguez[2], David Perea[3]
[1]Porto Business School, Porto University, Portugal, [2]Department of Financial Economics and Operation Management, University of Seville, Spain, [3]Department of Economics, Management, Industrial Engineering and Tourism, University of Aveiro, Portugal and Department of Financial Economics, Accounting and Operations Management, University of Huelva, Spain.

Abstract

This paper analyses the content of accommodation´ profiles on the social network Instagram, by applying an engagement analysis, to deepen in the scientific knowledge of the Instagram's role in the tourism industry. In this context, the authors aim to examine how the co-creation of experiences between accommodation providers and tourists influences the engagement of the latter associated with destinations. This study has significant managerial implications as it looks at how tourist companies can use social media to promote memorable tourist experiences that influence satisfaction and loyalty towards a destination. The innovation of this research is based on the methodology applied and data obtained in the tourism sector. An OSINT technique is applied to perform web scraping on Instagram capable of obtaining all the posts (n=10,017) of the profiles with their associated metadata. By analyzing their content and engagement, accommodation providers can develop more effective strategies to improve their brands. The article discusses how tourists use social media to express their perceptions towards tourist service brands or destinations. Finally, the study highlights the relationships between the engagement of tourists towards a destination with the content category and the media type on Instagram as they are crucial in shaping future tourists' perceptions.

Keywords: *Tourism experiences; Co-creation of experiences; Engagement analysis; Social media; OSINT*

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16429

Georeferencing sentiment scores to map and explore tourist points of interest

Luigi Celardo[1], Michelangelo Misuraca[2], Maria Spano[3]
[1]Department of Social Science, University of Naples Federico II, Italy, [2]Department of Business Administration and Law, University of Calabria, Italy, [3]Department of Economics and Statistics, University of Naples Federico II, Italy.

Abstract

Tourists are increasingly involved in co-creating attractions' symbolic images, sharing their experiences and opinions on websites like TripAdvisor and other similar rating and review platforms. In this paper, we propose a strategy for analyzing people' opinions about tourist points of interest, using an Ambient Geographic Information approach to georeference the polarity scores of reviews. Visualizing these scores on a map can be used to obtain helpful information for implementing strategic actions and policies of institutional and business actors involved in the tourist industry, as well as to help users plan their future experiences. A case study concerning the reviews of the restaurants in Naples (Italy) shows the effectiveness of the proposal.

Keywords: *social media, polarity score, georeferenced sentiment.*

1. Introduction

The diffusion of Web 2.0 changed in a few years each aspect of everyday life related to social representation and interaction. In a more flexible and disintermediated society, we are observing a growing consideration of how people – increasingly embedded in a digital ecosystem – communicate their feelings, opinions and experiences. This behavior is part of our reality, in accordance with the online and offline interconnectedness posited by Jurgenson (2019). This view implies a new communication paradigm in the definition of an experience of buying and consuming a product or service, including tourism-related ones. Today, social media guesting the narratives of people's experiences are used by the diverse actors involved in the tourist industry. Visitors, in particular, play a key role in co-creating tourist attractions' symbolic image, and the reputation generated by digital platforms is becoming progressively significant. Many individuals rely heavily on reviews and free-text comments on platforms such as TripAdvisor, Booking or Google when choosing a destination, reserving accommodations or a table in a restaurant, visiting cultural, scenic or amusement attractions. On the other hand, the use of people's opinion in tourist industry can contribute positively to the decision-making processes of institutions and businesses in fostering sustainability and understanding consumer behavior patterns.

In this scenario, the so-called *electronic word-of-mouth* becomes a primary source of information that deserve to be taken into account (Cheung & Thadani, 2012). It is necessary to monitor what visitors say about their experiences and feelings (Nambisan & Watt, 2011). Analyzing people's comments and reviews is essential since dissatisfied visitors are potentially dangerous, triggering a vicious circle of lousy reputation (Shirdastian, Laroche, & Richard, 2019), deterioration of symbolic value and significant financial losses (Luo, 2009). Nevertheless, these comments are composed as free text, and the information is encoded in a form that is difficult to process automatically. In a text mining framework, it is possible to pre-treat a textual body made of these comments and transform the unstructured data into structured data, performing statistical analyses to extract and manage the underlying knowledge base. Among the different tasks of text mining useful in the tourist industry, a significant role is played by the detection of the semantic orientation of texts, expressing the so-called *sentiment* in a numerical form. The *sentiment* resulting from the reviews written by tourists for a particular activity or attraction can be interpreted as a quantitative feature of a process involving their narratives about the experience.

Several alternative approaches have been proposed to calculate a score representing texts' negative/positive orientation and employ these scores in an opinion-mining strategy (Hemmatian & Sohrabi, 2019). Contributions proposed in the literature have been more oriented towards topic extraction (Zhao *et al.*, 2016) or classification (Kim & Lee, 2014), where instead the visualization of sentiment is still an open research topic (Kurcher *et al.*, 2018).

In a tourist domain, the geographical dimension has to be wisely considered to better analyze its socio-economic traits. In this work, we propose a strategy based on the computation of polarity scores for a set of reviews related to some points of interest and their spatial localization. Georeferenced data, typically represented by a set of geographic coordinates, allows visualizing and understanding spatial patterns and relationships that may not be apparent from other types of data. Latitude and longitude coordinates may be then used to georeference the sentiment and visualize the semantic orientation of each tourist point of interest on a map. Our strategy relies on the *Ambient Geographic Information* (AGI) approach (Stefanidis *et al.*, 2013), in which social media are used to understand the human landscape and its evolution over time. Starting from the AGI framework, we extended the concept to digital platforms like TripAdvisor. Other authors considered the importance of using sentiment data and geographical references. The semantic orientation of comments posted on Twitter have been geographically analyzed to study the density of unfavorable/favorable opinions across U.S. (Camacho *et al.*, 2021). The joint use of sentiment and geographical data has been also used to assess the socio-environmental impact of large-scale infrastructure projects (Li et *al.*, 2021). In a tourist domain, the *tourism sustainability index* (TSI), a synthetic indicator encompassing a dimension based on polarity scores, has been proposed to frame and georeference tourist satisfaction in accordance with the European Tourism Indicator System (De Marchi *et al*., 2022).

From a theoretical viewpoint, the sentiment scores and other metadata may be used to cluster the points of interest in a spatial perspective, better guiding the actions of operators and institutional actors and people interested in exploring an area to plan a visit or a journey. In the following, a case study based on the city of Naples (Italy) is presented to show the proposal's effectiveness. Naples is nowadays one of the most important tourist attractions in Italy. Its historic center has been on the UNESCO World Heritage List since 1995 but has only become a primary destination in the last few years. Here, we are particularly interested in evaluating restaurants because Neapolitan restaurants are considered part of the tourist experience (Vrontis *et al.*, 2021), as the city is famous for its cuisine and gastronomic heritage. Furthermore, the food & wine supply chain is an essential driver of the local economy, creating jobs and promoting the area (Della Corte *et al.*, 2015).

2. Materials and methods

The restaurant reviews used in this study have been scraped from the Italian TripAdvisor website, using `Naples` in the query and `restaurants` as the main category, and stored in a local repository together with some metadata: *name*, *address*, *latitude* and *longitude* (validated with the corresponding Google Maps `id place`), *rating*, *# of reviews*. At the current stage, we considered 774 activities (a share of 30% with respect to the total number

of 2,634 restaurants in Naples). Moreover, we decided to set a limit of 1,000 reviews per restaurant – namely the most recent ones – obtaining a collection of 283,801 reviews.

To perform a lexicon-based sentiment analysis of this collection, we used an original customized lexicon of Italian terms. Most resources in the sentiment research area, like lexicons, labelled collections and NLP tools, are mainly available in English. The lack of linguistic resources is critical in the majority of studies, producing a so-called *lexical gap* (Chiavetta *et al.*, 2016). Thus, we built an Italian lexicon by merging the resources developed in the Sentix project (Basile & Nissim, 2013), the Opener project (Russo *et al.*, 2016), and other aptly screened studies (e.g., Bolasco & Della Ratta, 2004). The resulting lexicon contains 26,511 polarized terms with a value of +1 if positives and -1 if negatives.

For our analysis, a light pre-treatment was applied to the reviews. Non-alphabetic characters and symbols – like numbers or emoticons – were removed to consider only content-bearing words. The polarity scores have been calculated with a sentence-level logic (Balbi *et al.*, 2018). Each review is segmented into its constituent sentences to consider the sentiment associated with each aspect of the described experience. Given a review r_i (i=1,…,n), its a_i sentences $\{s_{i1}, ..., s_{ik}, ..., s_{ia_i}\}$ are identified by considering as separators strong punctuation marks like full stops, question marks and exclamation marks. The k-th sentence s_{ik} is represented as a sequence of its p_k terms $\{w_{ik1}, ..., w_{ikj}, ..., w_{ikp_k}\}$. Each term w_{ikj} in the k-th sentence of the i-th review is compared with the terms in the lexicon, assigning -1 to negative terms and +1 to positive terms, respectively. Terms not listed in the lexicon are scored with a null value. The polarity of each term is then weighted considering negators (e.g., *mai*, *nessuno, nessuna*), amplifiers and de-amplifiers (e.g., *poco, molto, pochissimo*), adversative and contrasting terms (e.g., *ma*, *tuttavia*). This weighting scheme allows for emphasizing or dampening the negativity or positivity of each polarized term, leading to a more effective measure of semantic orientation (Vechtomova, 2017). The polarity score of each sentence $\mathrm{PS}_{s_{ik}}$ is obtained as the sum of weighted term scores $\mathrm{PS}_{w_{ikj}}$ on the sentence length:

$$\mathrm{PS}_{s_{ik}} = \frac{\sum_{j=1}^{p_k} \mathrm{PS}_{w_{ikj}}}{\sqrt{p_k}} \tag{1}$$

Since we are interested in obtaining a polarity score at a review level, we calculated an overall score PS_{r_i} for each text by a down-weighted zeros average of sentence polarities, giving a lower weight to sentences conveying a neutral sentiment:

$$\mathrm{PS}_{r_i} = \frac{\sum_{k=1}^{a_i} \mathrm{PS}_{s_{ik}}}{a_i^- + a_i^+ + \sqrt{\log(1+a_i^0)}} \tag{2}$$

where a_i^-, a_i^+ and a_i^0 are the numbers of sentences in r_i with a negative, positive, or neutral polarity, respectively. Figure 1 graphically depicts an example of the polarity score computation, reporting for each review the overall score together with negative aspects (in red) and positive aspects (in green) of the tourist experience.

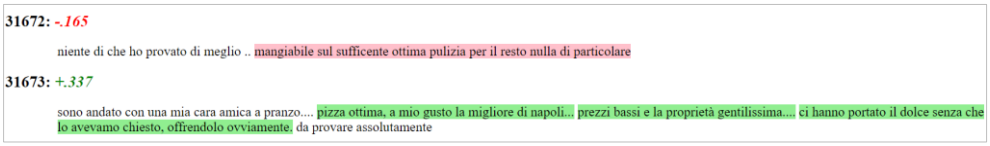


Figure 1. Examples of reviews with polarity score and highlighted negative/positive aspects.

Since the scores PS_{r_i} assume values in a]-∞,+∞[interval, we decided to rescale all the results in a]0,1[interval to facilitate the interpretation. In the subsequent step of the strategy, sentiment scores have been used to characterize each restaurant. Specifically, we used latitude and longitude to plot the restaurants on a map and visualize the results of the sentiment analysis as hot spots. We categorized the sentiment scores into low ($PS_{r_i} < 0.3$), medium ($0.3 \leq PS_{r_i} \leq 0.6$), and high ($PS_{r_i} > 0.6$), using a red-to-green color palette for the gradient. The different actors can use the resulting representation to explore the specific area under investigation, whereas researchers can include this information in more articulated analytic strategies. Georeferenced polarity scores can be used with other metadata to cluster points of interest identifying groups of activities that share similar characteristics or attributes. This can be particularly valuable both from an urban planning and tourist marketing side, where can be crucial to understand the characteristics of different neighborhoods or areas of a city and develop targeted strategies or interventions.

3. Some preliminary results

In Figure 2, we can see the polarity scores of Neapolitan restaurants georeferenced on the city map. The output obtained by applying the strategy can be interactively browsed by zooming on the different points. Here we used a static screenshot by way of illustration. Green areas represent the restaurants with a higher positive sentiment, whereas red areas represent the restaurants with a higher negative sentiment. The map shows a concentration of restaurants with a positive sentiment near the city's historical center and the waterfront near the so-called Riviera di Chiaia, rich in tourist attractions and gastronomic sites. The hot areas, associated with the red color, are mainly located in the peripheral or ex-industrial districts, such as Bagnoli (on the left side of the map) or San Giovanni a Teduccio (on the right side of the map). Although these districts of Naples have tourist potential (e.g.,

Bagnoli hosts the Science Museum, San Giovanni hosts the National Railway Museum), they have critical deficiencies in the sphere of transport, services and air quality.

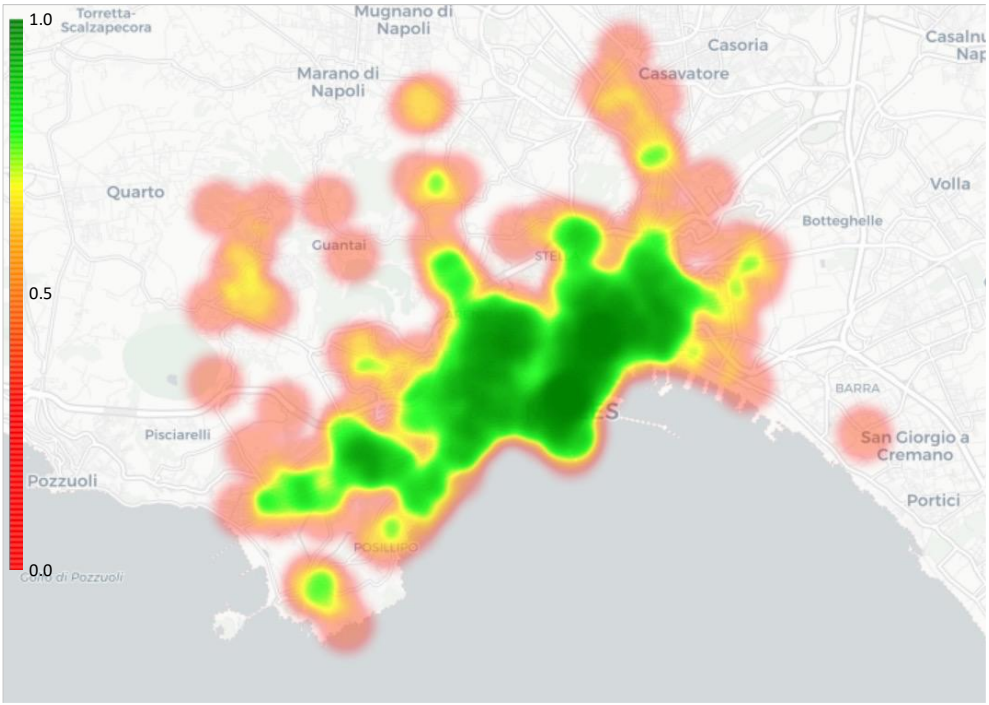


Figure 2. Visualization of restaurants' sentiment in the city of Naples.

According to a study carried out by La Rocca (2021) on the city of Naples, the poor accessibility of peripheral areas (in terms of public transport) and the degradation of urban spaces are, among others, the primary grounds for complaints, affecting negatively the intention of visitors to come back again to the city. Restaurants and shops generally suffer a lack of infrastructure and services (Buonanno *et al.*, 2009). Moreover, the absence of action to protect and enhance the historical and cultural heritage can reduce tourist attractiveness, as in the case of the San Giovanni district.

The complete results of the case study will be discussed more in detail elsewhere.

References

Balbi, S., Misuraca, M. & Scepi, G. (2018). Combining different evaluation systems on social media for measuring user satisfaction. *Information Processing & Management*, *54*(4), 674-685.

Basile, V., & Nissim, M. (2013). Sentiment analysis on Italian tweets. In A. Balahur, E. van der Goot & A. Montoyo (Eds.), *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 100-107). Association for Computational Linguistics.

Bolasco, S., & Della Ratta, F. (2004). Experiments on semantic categorisation of texts: analysis of positive and negative dimension. In G. Purnelle, C. Fairon & A. Dister (Eds.), *Le poid des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles* (pp. 202-210). UCL Presses Universitaires de Louvain.

Buonanno, G., Morawska, L., Stabile, L., & Viola, A. (2009). PM concentrations in pizzerias: implications for occupational exposure. *Journal of hazardous materials*, *164*(2-3), 870-874.

Camacho, K., Portelli, R., Shortridge, A., & Takahashi, B. (2021) Sentiment mapping: point pattern analysis of sentiment classified Twitter data. *Cartography and Geographic Information Science*, *48*(3), 241-257.

Cheung, C.M.K., & Thadani, D.R. (2012). The Impact of Electronic Word-of-Mouth Communication: A Literature Analysis and Integrative Model. *Decision Support Systems*, *54*, 461-470.

Chiavetta, F., Lo Bosco, G., & Pilato, G. (2016). A Lexicon-based Approach for Sentiment Classification of Amazon Books Reviews in Italian Language. In T.A. Majchrzak, P. Traverso, V. Monfort, K.-H. Krempels (Eds.) *Proceedings of the 12th International Conference on Web Information Systems and Technologies* (pp. 159-170). Scitepress.

Della Corte, V., Sciarelli, M., Cascella, C., & Del Gaudio, G. (2015). Customer Satisfaction in Tourist Destination: The Case of Tourism Offer in the City of Naples. *Journal of Investment and Management*, *4*(1-1), 39-50.

De Marchi, D., Becarelli, R., & Di Sarli L. (2022). Tourism Sustainability Index: Measuring Tourism Sustainability Based on the ETIS Toolkit, by Exploring Tourist Satisfaction via Sentiment Analysis. *Sustainability*, *14*(13), 8049-8067.

Hemmatian, F., & Sohrabi, M.K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, *52*, 1495-1545.

Jurgenson, N. (2019). *The Social Photo: On Photography and Social Media*. New York, NY: Verso Books.

Kim, K., & Lee, J. (2014). Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. *Pattern Recognition*, *47*, 758-768.

Kurcher, K., Paradis, C., & Keren, A. (2018). The state of the art in sentiment visualisation. *Computer Graphics Forum*, *37*, 71–96.

La Rocca, R.A. (2021). Urban Accessibility and Tourist Activity: An Application to the Metropolitan City of Naples. In D. La Rosa, & R. Privitera (Eds.), *Innovation in Urban and Regional Planning - INPUT 2021* (pp. 583-591) Springer.

Li, Y., Zhang, Y., Tiffany, L. A., Chen, R., Meng, C., & Liu, J. (2021). Synthesizing social and environmental sensing to monitor the impact of large-scale infrastructure development. *Environmental Science & Policy*, *124*, 527-540.

Luo, X. (2009). Quantifying the Long-Term Impact of Negative Word of Mouth on Cash Flows and Stock Prices. *Marketing Science*, *28*(1), 148-165.

Nambisan, P., & Watt, J.H. (2011). Managing customer experiences in online product communities. *Journal of Business Research*, *64*, 889–895.

Russo, I., Frontini, F., & Quochi, V. (2016). *OpeNER Sentiment Lexicon Italian - LMF*, Institute for Computational Linguistics "A. Zampolli", National Research Council, Pisa. http://hdl.handle.net/20.500.11752/ILC-73.

Shirdastian, H., Laroche, M., & Richard, M.-O. (2019). Using big data analytics to study brand authenticity sentiments. The case of Starbucks on Twitter. *International Journal of Information Management*, *48*, 291-307.

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, *78*(2), 319–338.

Vechtomova, O. (2017). Disambiguating context-dependent polarity of words: An information retrieval approach. *Information Processing & Management*, *53*, 1062-1079.

Vrontis, D., Basile, G., Tani, M. & Thrassou, A. (2021). Culinary attributes and technological utilization as drivers of place authenticity and branding: the case of Vascitour, Naples, *Journal of Place Management and Development*, *14*(1), 5-18

Zhao, Y., Qin, B., Liu, T., & Tang, D. (2016). Social sentiment sensor: A visualization system for topic detection and topic sentiment analysis on microblog. *Multimedia Tools and Applications*, *75*(15), 8843-8860.

How can destinations get engagement on Instagram? Artificial Intelligence as a tool for photo analysis

Sofía Blanco-Moreno[1], Ana M. González-Fernández[1], Pablo Antonio Muñoz-Gallego[2]
[1]Business Management and Economics Department, University of León, Spain, [2]Business Administration and Economics Department, University of Salamanca, Spain.

Abstract

What type of content should be published on Instagram to get more engagement? This article highlights the different characteristics that the images of tourists show on Instagram with the most engagement, that is likes and comments. Understanding the behavior in a destination helps tourism managers in marketing strategies. Based on the stimulus-organism-response model, a content analysis of 49,540 photographs shared by tourists that received 3,734,384 likes and 133,497 comments is carried out. By combining the content analysis with Kruskal-Wallis non-parametric tests, the results show that the different characteristics found in the images imply different amounts between comments and likes, demonstrating that the behavior of users on Instagram is influenced by the different attributes of the images. Specifically, images that feature people get more engagement than destination-focused ones. Additionally, scenes such as gastronomy and nature get less engagement than scenes such as old and new heritage, outdoors, and entertainment. Specifically, photos with people get greater rate of comments than likes, and if the format is selfie, they also get more comments. The implications of this research directly affect destination managers, offering clues about the content generated by tourists that produces the most engagement, thus attracting potential tourists and Instagram users.

Keywords: *Destination image; SOR model; Instagram; Visual computing, Selfie; Neural networks.*

UGCs and wellness touristic image: the Spanish case

Myriam González-Limón[1], Lourdes Cauzo-Bottala[2], Rocío Martínez-Torres[2], F. Javier Quirós-Tomás[2]

[1]Departamento de Análisis Económico y Economía Política, Universidad de Sevilla, Spain,
[2]Departamento de Administración de Empresas y Marketing, Universidad de Sevilla, Spain,

Abstract

The purpose of this paper is to analyse the characteristics of the projected image of wellness tourism by studying memorable experiences transmitted through user-generated content (UGC) in eight Spanish tourist destinations.

To achieve this objective the methodology employed has been a netnographic and framework analysis applied to a UGC dataset collected from Airbnb Experiences in eight Spanish tourist destinations. Based on the keyBERT value, the dimensions and elements that characterise wellness were identified, and a correlation analysis was carried out. Using these dimensions and the UGC of each destination, the wellness image of each tourist destination was identified. The main result is that the image of a tourist destination can be established on the basis of the UGC. From all wellness dimensions (Body, Spirit, Mind, Social and Environment), the Spirit dimension stands out as the most relevant in the image of the destination when we talk about wellness tourism. Likewise, the existence of strong linear correlations, both positive and negative, between the wellness dimensions and their elements is also observed.

The interest of the work lies in the use of data from sources that have been little exploited scientifically in order to test their validity as a source of projected tourist image of different destinations, applied to wellness tourism. It seeks to confirm the validity of the set of keywords found in order to create a valid library for future studies on wellbeing based on UGC analysis.

Keywords: *UGC, destination image, tourist image, wellness tourism, experience tourism.*

This work was supported by the project "Identificación de los Atributos Únicos de los Destinos Turísticos Andaluces desde la perspectiva de los Social Media mediante el uso de técnicas de Text Mining (TURIMEDIA)" P20_00639 funded by Junta de Andalucía.

Not your fault, but your responsibility: worsened consumer sentiment on work-from-home products during COVID-19

Giovanni Cintra[1] , Filipe Grilo [1,2,3]
[1] University of Porto, School of Economics and Management, Portugal [2] CEF.UP, Portugal
[3] CEAD, Portugal.

Abstract

This study analyses the evolution of people's sentiment towards Work from Home (WFH)-related products during the pandemic, using user-generated content from Twitter on responses for the largest US online furniture stores.

The goal of this study is threefold. First, we test if Covid-19 disrupted the volume of Electronic Word of Mouth for WFH-related products and if Covid-19 changed people's sentiment toward WFH-related products. Finally, we assess which online furniture stores had a more positive or negative impact on sentiment during the covid-19 outbreak.

We find that people interacted more about WFH products during the Covid-19 lockdowns, but sentiment towards WFH products worsened. For some online furniture stores, Covid-19 restrictions may explain the changes in sentiment, but firms' idiosyncrasies also play a role.

The methodology of this study allows companies to assess the impact of external effects on customers' sentiments, allowing them to identify specific problems and to connect more naturally with their customers.

Keywords: *Covid-19, Electronic Word of Mouth, Sentiment Analysis, Twitter, Work From Home.*

Corresponding author. E-mail address: fgrilo@fep.up.pt.

Acknowledgments: Filipe Grilo acknowledges that his research has been financed by Portuguese public funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., in the framework of the project UIDB/04105/2020.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16432

Assessing the spread of Keynesian ideas in the economic policy debate: a Text Mining approach on Twitter

Chiara Perfetto[1], Antonella Rancan[1], Giuliano Resce[1]
[1] Department of Economics, University of Molise, Italy

Abstract

This paper proposes a methodology for examining the presence of Keynesian ideas in the economic debate. To this aim we use Twitter as a source of data to monitor the debate in real time. We quantify the presence of Keynesian and anti-Keynesian thought in tweets about the economy and we qualify the emotional tone of these tweets. Our preliminary results show that the 20 per-cent of total English tweets about #economy contain words related to Keynes while about 8 per- cent contain words referring to anti-Keynesian policies. The monthly analysis of the tweets shows a certain heterogeneity. The distribution of Keynes-related tweets is much more uneven than the distribution of anti-Keynesian tweets. Our evidence suggests that the methodology we applied to understand how much of the Keynesian thought is still around in the economic debate can be promising. The next step will be to focus on georeferenced tweets to detect heterogenity across countries and to understand how country-level trends reflect the economy cycle. This study still has some limitations that will be faced in future research such as the classification of topics and the focus on English texts for the moment.

Keywords: *Text Mining; Twitter; Keynesian thought.*

1. Introduction

Social media interactions, from commenting on a post to liking a photo, leave digital traces that can be used to extract patterns of individual, group and social behaviours. In this paper we address how social data created by users are useful indicators for providing insights about economic patterns. The political science literature shows how the feedback mechanism reinforces the entrenchment of existing policies in the case of 'positive' feedbacks or subvert current patterns due to 'negative' feedbacks (Pierson & Skocpol, 2002).

Social scientists are often unable to systematically assess the significance and impact of ideas and research outputs due to extensive research portfolios, time and resource constrains. Therefore, the dynamics of knowledge dissemination are not fully understood. The integration of digital research methodologies, with text mining techniques offers an innovative and comprehensive approach to deal with such challenges. In this study we suggest Twitter as a source to study how public opinion relies on Keynesian or 'anti-Keynesian' view when debating economic issues. Data from Twitter make it possible to monitor the onset and spread of phenomena in real time (Resce & Maynard 2018). In fact tweets have a reliable timestamp and for that they can be analyzed from a time perspective and are accessible to researchers, unlike most social networking sites (Fujiwara et al., 2021). For these reasons, Twitter has found many applications among social scientists for many different purposes as detecting tourism preferences (Chang, Chu, 2013), analysing political trends (Rill et al., 2014, Seabold et al., 2015), or studying socio-economic problems (Resce, Maynard, 2018).

With this research, we aim to identify a methodology to quantify the presence of Keynesian and anti-Keynesian ideas in tweets about the economy and to qualify the emotional tone of these tweets (Misuraca et al., 2020). In this paper the terms "Keynesian" and "anti-Keynesian" are broadly defined. For Keynesian thought or Keynesian view we mean an approach to economic problems such as unemployment and economic downturn which relies on public interventions for their overcoming. On the contrary, anti-Keynesian view is defined in terms of a free-market approach, as supported by neoclassical economics. Twitters which mention government spending policies and expansionary monetary actions are Keynesian, while restrictive, anti-inflationary policies and deregulation policies are considered anti-Keynesian.

2. Data and method

The inclusion of Twitter data to understand the economic trend involves web scraping techniques on existing official Twitter accounts. The scraping was based on preliminary criteria which come from searching of a specific hashtag. We downloaded 7.255.518 tweets from 2008 to March 2022 regarding the hashtag of interest, that is #economy. For every

tweet, the following fields have been considered: URL of the tweet; text of the tweet (along with the extraction of hashtags and mentions); timestamp, i.e., time of tweet creation; username of the publisher; number of likes; number of retweets; number of replies; geolocation (if available); language.

To quantify how many Tweets reflect Keynesian or anti-Keynesian thought we built a taxonomy taking into consideration a subsample of tweets containing the word "Keynes". From this subsample of 3,253 tweets we have extracted single words, bigrams, and trigrams which have a frequency higher than 3% of the tweets containing at least one Keyn-based word. For the extraction of the single words a Term Document Matrix was produced, with tweets id by column and stemmed words by row. The Term Document Matrix indicates the number of times each word appears in each tweet. For the bi-grams and tri-grams we used the function to tokenize in consecutive sequences of words, called n-grams. As one might expect, a lot of the most common bi-grams are “stop-words”, as "of the", "to be", etc. For that we removed cases where one or two of the two topics is a stop-word. The same was done for the tri-grams. Then the n-grams extracted were manually labelled (a strategy used by recent studies, such as Angelico et al., 2022) as Keynesian policy related or anti-Keynes policy related as in the Table1. Unigram, bigram and trigram where used, in the case of anti-Keynesian n-grams trigram where not connected to the topic and this is the reason why they were not considered.

Table 1. Taxonomy

PRO KEYNES			ANTI KEYNES		
TOPIC	BIGRAM	TRIGRAM	TOPIC	BIGRAM	TRIGRAM
Keynesian	maynard keynes	john maynard keynes	hayek	neoclassical economics	
Keyn	john maynard	wethepeople economy obama	auster	miltonfriedman rocks	
Invest	keynesian economics	economics rukvtbw2 dems	inflat	money economy	
Obama	keynesian economic	proof deficit spending	right	hayek round	
Stimulus	keynesian economists	feed finance economy	neoclass	austrian economy	
Recess	economy keynes	dems econ economy	teaparti	hayek rap	
Debt	keynes economics	maynard keynes economy	friedman	tcpt teaparty	
Deficit	keynes economy	maynard keynes considered	trump	modern economics	
Johnmaynardkeyn	keynesian economy	investing economy guest	miltonfriedman	milton friedman	
Nyt	economy keynesian	keynesian economics economy	neoliber	economy inflation	
economic	investing economy		republican	economy austerity	
Keynesianeconom	paul krugman		libertarian	keynesian fail	
Marx	economy obama		reagan	political economy	
Marxism	deficit spending		freemarket	ron paul	
Skidelski	rocks keynesian		monetarist	anti keynesian	
Democrat	keynesian proof			zerohedge kenesian	
Nytim	keynesian stimulus			economy teaparty	
	growing economy				
	economic growth				
	proof deficit				
	keynesian theory				
	economix blog				
	keynesian policies				
	capitalism money				
	keynes bit				
	century keynes				
	economy jobs				
	economy recession				
	obama economy				
	economic recovery				
	economy debt				
	government economy				
	government spending				
	keynes considered				
	post keynesian				
	neo keynesian				
	keynesian model				
	economy investing				
	keynesianism economy				
	nyt economix				
	economy keynesianeconomics				
	keynesian depression				
	cambridge corridor				
	debt economy				
	economics keynes				
	economy stimulus				
	jm keynes				
	keynesian economist				
	keynesian money				
	keynesian multipler				
	krugman keynes				
	obama kenesian				
	robert skidelsky				
	socialism marxism				
	stimulus economy				
	general theory				

To quantify what each tweet is about and if the tweet is connected to Keynesian or anti-Keynesian thought and policy, we used text mining techniques. For text mining we reduced the dataset to only 6.457.704 English tweets to facilitate some functions such as stemming and sentiment. Using R language, we adopted extensive customization of existing tools and algorithms to conduct such analyses. The text corpus of analysis was prepared using functions from the R package "tm" (Feinerer and Hornik, 2018; Feinerer, Hornik, and Meyer, 2008): punctuation, stop words (i.e., in English, words like "the", "is", "of", etc), special characters and numbers were removed from the corpus. The words were then converted to lowercase and stemmed. Also, the topics in the taxonomy were converted to lowercase and stemmed.

The words of the taxonomy are identified and counted in the tweets through the functions of the "stringr" package on R (Wickham, 2019). Keynes related tweets are defined as the sum of tweets featuring one or more words contained in the list of Keynesian topics while tweets related to anti-Keyensian policies result from the sum of tweets that contain one or more words contained in the list of anti-Keynesian topics.

To identify the semantic orientation of each line of text downloaded from Twitter we have adapted the functions of the "sentimentr" package developed by Rinker (2017) which uses a dictionary-based approach, i.e. based on a predefined polarized word list. Once the sentiment has been estimated, we produce graphs capable of highlighting the trend of sentiment over time. The set of information obtained by the text mining plus the number of tweets combined with the sentiment could be combined to build a sentiment index.

3. Preliminary results

The 19.3% of the economy total English tweets contain words related to Keynesian ideas while 7.7% contain words referring to an anti-Keynesian policies. The monthly analysis of the tweets (see Figure 1) shows a greater homogeneity of the distribution of anti-Keynesian tweets up to July 2016 with a subsequent increase in the volume of Tweets until reaching the peak in March 2019.

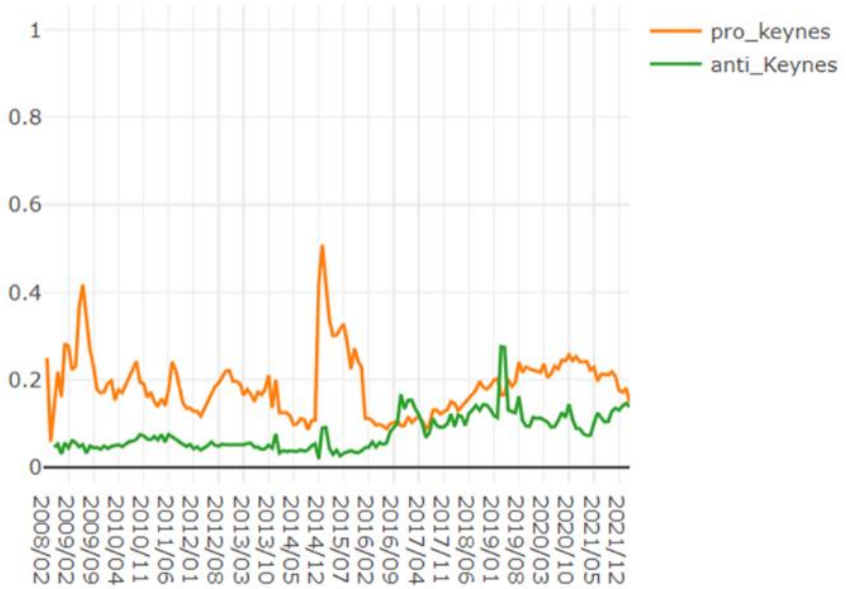


Figure 1. Distribution of tweets in timeline

As we can see from Figure 2 (top left) the distribution of Keynes-related tweets is much more uneven than the distribution of anti-keynesian tweets. The first peak is reached in June 2009, when 40% of the economic tweets written in that period contain words referring to Keynes. The largest volume of tweets referring to Keynes is reached in January 2015. This may be connected to the desire to abandon neoclassical theories following the 2007/2008 financial crisis and the subsequent Great Recession. If we look at the remaining part of Figure 2, we can also see an irregular distribution of likes, retweets and replies.

As we can see from Figure 3, the increase in the volume of anti-Keynesian tweets after 2016 corresponds to an increase in likes and retweets while the most replies tweets are those of 2009 and 2010.

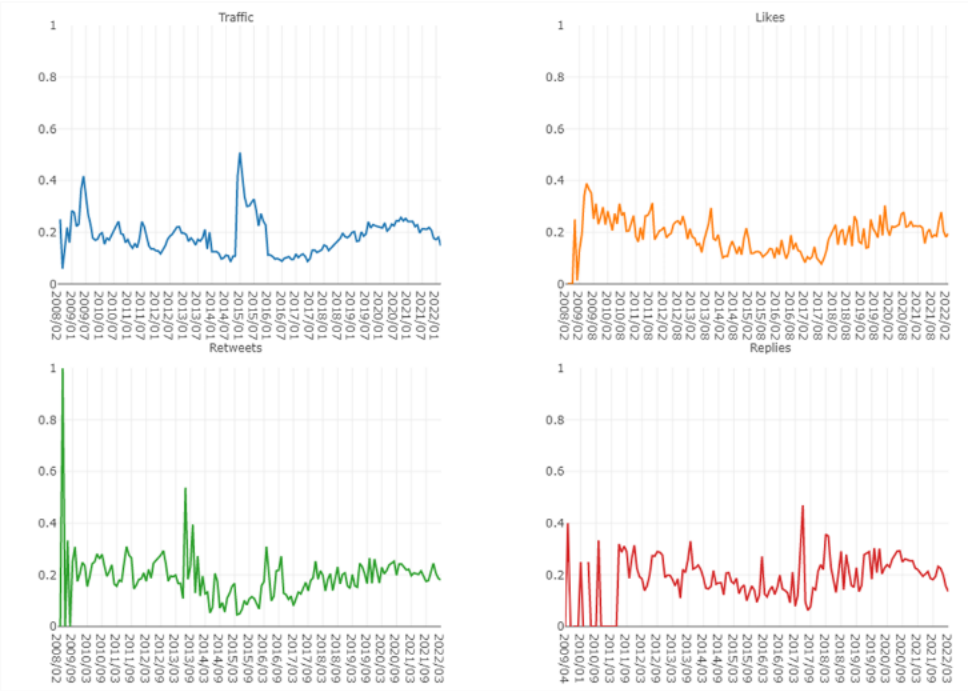


Figure 2. Statistics on Keynes-related tweets

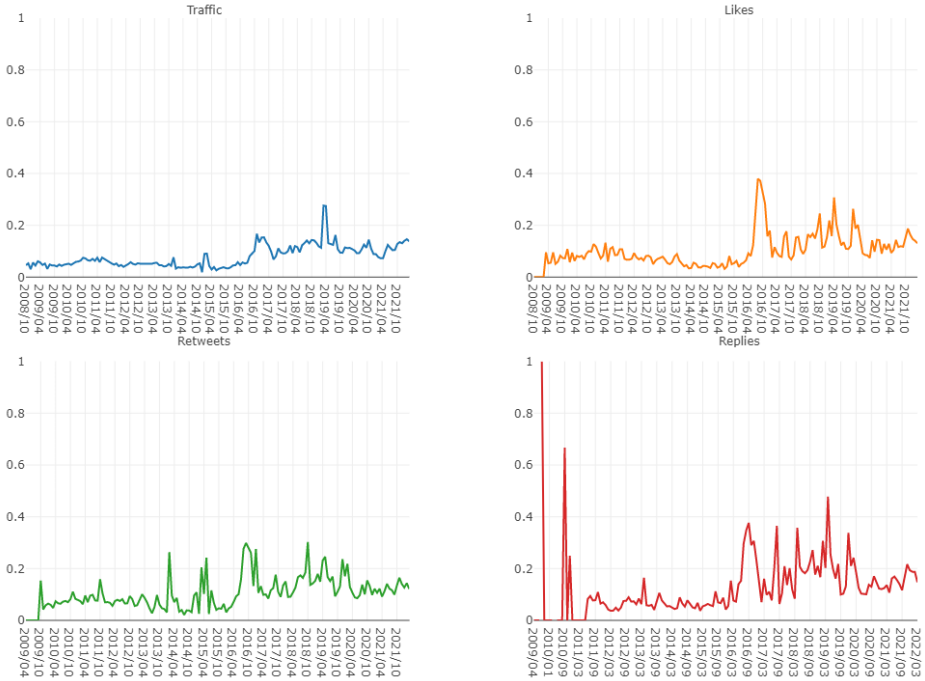


Figure 3. Statistics on antiKeynes-related tweets

As in the Figure 4, the sentiment relating to anti-Keynesian tweets is almost positive. Keynes-related tweets are associated with more volatile sentiment through 2014. Post-2014 the average sentiment of tweets is positive. Combining the results from supervised text mining with the sentiment generates an index that allows to compare the average sentiment trend of Keynesian or anti-Keynesian tweets (Figure 5).

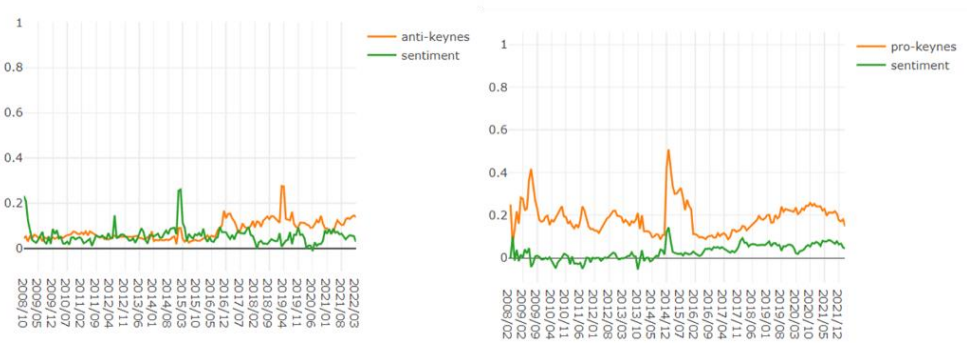


Figure 4. Sentiment analysis trends

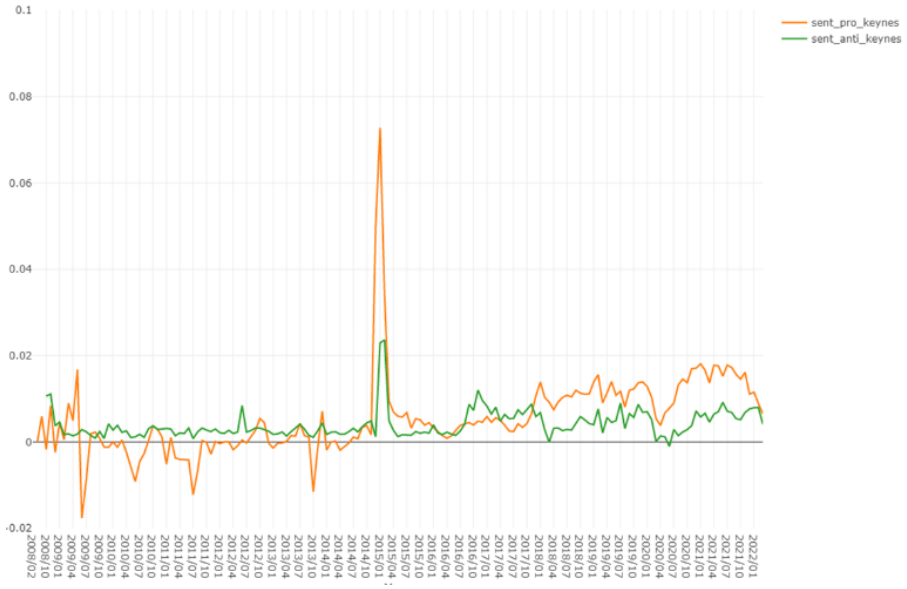


Figure 5. Sentiment index

The preliminary evidence of this research suggest a methodology for an understanding of the main references to which the economic policy debate carried out throug social media relies on. In line with previous literature, our results suggest that the integration of text mining techniques with Twitter data has a great potential to add knowledge to social science (Chang, Chu, 2013; Rill et al., 2014; Seabold et al., 2015; Resce, Maynard, 2018). For our future research directions, we plan to reduce the set of tweets at only georeferenced tweets to focus these trends across countries and understand how these trends reflect the economic cycle.

This study also includes some limitations that should be taken into account. When designing a study based on the taxonomy, there may be subjectivity in the classification of topics. Also the reduction to English texts only, necessary above all for the sentiment, could generate loss of information from the general dataset.

References

Angelico, C., Marcucci, J., Miccoli, M., & Quarta, F. (2022). Can we measure inflation expectations using Twitter?. *Journal of Econometrics*, 228(2), 259-277.

Chang, C. C., & Chu, K. H. (2013, June). A recommender system combining social networks for tourist attractions. In 2013 Fifth International Conference on Computational Intelligence, Communication Systems and Networks (pp. 42-47). IEEE.

Feinerer, I. (2008). An introduction to text mining in R. R News, 8(2), 19-22.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of statistical software*, 25(5), 1-54.

Fujiwara, T., Müller, K., & Schwarz, C. (2021). The effect of social media on elections: Evidence from the United States (No. w28849). *National Bureau of Economic Research*.

Misuraca, M., Forciniti, A., Scepi, G., & Spano, M. (2020). Sentiment Analysis for Education with R: packages, methods and practical applications. arXiv preprint arXiv:2005.12840.

Pierson, P., & Skocpol, T. (2002). Historical institutionalism in contemporary political science. *Political science: The state of the discipline*, 3(1), 1-32.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89, 14-46.

Resce, G., & Maynard, D. (2018). What matters most to people around the world? Retrieving Better Life Index priorities on Twitter. *Technological Forecasting and Social Change*, 137, 61-75.

Rill, S., Reinel, D., Scheidt, J., & Zicari, R. V. (2014). Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. Knowledge-Based Systems, 69, 24-33.

Rinker, T. (2017). sentimentr: Calculate Text Polarity Sentiment. R package version 2.9.0.

Seabold, S., Rutherford, A., De Backer, O., & Coppola, A. (2015). The pulse of public opinion: using Twitter data to analyze public perception of reform in El Salvador. World Bank Policy Research Working Paper, (7399).

Wickham, H. (2019). stringr: Simple, consistent wrappers for common string operations. R package version 1.4. 0.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16474

Measuring Social Mood on Economy during Covid times: effects of retraining Supervised Deep Neural Networks

Elena Catanese[1], Mauro Bruno[1], Luca Stefanelli[2], Francesco Pugliese[1]
[1]Istat, Rome Italy [2]University of Bergamo, Italy

Abstract

Supervised Machine learning approaches are popular techniques used for sentiment analysis tasks. However, such techniques have strong limitations due to their sensitivity to the quantity and quality of the training datasets and may fail when training data are biased or insufficient. In the present study we address the impact of Covid on a deep learning classifier based on long-short term memory neural network (LSTM). This classifier is used to compute a daily sentiment index on Italian tweets with economic content, for the first five months of 2020 (more than 11 million of tweets are classified). We show how retraining the model with a set of annotated tweets containing reference to Covid increase the accuracy of the classifier. The accuracy is measured by analyzing the dynamics of the index. We will show that during pandemic the retrained index decreases coherently with most Italian economic indicators.In addition, we analyze how the training and tuning procedures (one-step, two-steps with fine-tuning) affect the daily dynamics of the index.

Keywords: *Sentiment Analysis, Artificial Neural Networks, Deep learning, Twitter data, Word Embedding Models*

1. Introduction

In this paper, we describe a new methodology to calculate a daily sentiment index that aims to depict the population mood about the economy in Italy, such as Social Mood on Economy Index (SMEI) (Catanese, et al., 2022). So far, Italian National Institute of Statistics (Istat) adopted an unsupervised procedure for estimating the social mood on economy by exploiting a lexicon-based approach (since 2018). Our initial choice fell on lexicons, due to the lack of Italian labelled datasets of tweets for sentiment analysis. Within the last few years, the number of labelled Italian datasets has increased, so we started to investigate cutting-edge supervised deep learning algorithms for sentiment analysis purposes. This work relies on a binary sentiment classifier (positive vs negative content) built by means of a Bidirectional Long short-term memory (LSTM) neural network. The training phase consists of a two-steps procedure: a) an unsupervised word embedding training model built on a unlabeled training set of Italian tweets extracted from the SMEI; b) a supervised training aimed at generating the model that computes the sentiment score.

Recently the outbreak of Covid changed structurally the content of twitter conversations. For this reason, we decided to retrain the original and previous neural network so to include a Covid related annotated dataset and to analyze accuracy improvements. In addition, we analyzed the impact on the daily index of the training procedure by splitting it into two steps. As a result, we analyzed four scenarios: original model; retrained models (both one step;two).

In this work we want to address the following question: how is the algorithm able to intercept Covid "drop" in social mood during the lockdown period.

The paper is structured as follows: in section 2 we describe LSTM methods, the different scenarios and the dataset used in our simulations. In section 3 the dynamics of the daily social mood index, within the period January 2020 - May 2020, according to the different scenarios are commented. Conclusions are drawn in Section 4.

2. Methods

2.1 Recurrent Model for Text Classification

Our classification model is a long short-term memory (LSTMs) (Hochreiter, & Schmidhuber, 1997) which are a type of Recurrent Neural Network (RNN) (Medsker, & Jain, 2001) able to process long sequences of data. In general terms, a LSTM memory cell is composed 4 gates: an input gate, an output gate, a forget gate and a self-recurrent neuro. The input and output gates control the interactions between neighboring memory cells and the memory cell itself. Whether the input signal can alter the state of the memory cell is controlled by the input gate. On the other hand, the output gate can control the state of the memory cell on whether it can alter the state of other memory cell. In addition, the forget gate can choose to remember or

forget its previous state. LSTMs layers are composed of a number of cells, usually activated by means of hyperbolic tangents, which regulate the flow of information through the system of gates, in the form of sigmoid functions. RNNs are suitable for Natural Language Processing (NLP) applications thanks to their capability to connect previous pieces of information to the present tasks. In this work, the proposed model is a neural network built upon a two-layer Bidirectional LSTM (Huang, et. al., 2015). Bidirectional LSTMs (BiLSTM) are an extension of LSTMs which can improve the performance depending on the given task. BiLSTMs are basically two independent LSTMs trained on the input sequence. The first LSTM processes the sequence data forwards, whereas the second LSTM processes the sequence data backward with two separate hidden layers. BiLSTM connects the two hidden layers to an output layer. This structure provides the network both forward and backward information at each step.

The key feature of BiLSTM is the ability to capture the "context" of each word within the text in a very powerful way since one LSTM works on the left context and the other LSTM "understands" the right context. For instance, the word "bank" can assume different meanings according to its context, the left context can provide one meaning such as in "river bank". But the right context matters too, because it provides another meaning, for instance in "Bank of America", the BiLSTM can capture these different gradients of meaning by means of the double LSTM.

2.2. Training Process and Computation of daily index

The input layer of the neural network is an embedding layer, i.e., an embedding space resulting from a word-embedding model. The embedding model has been built on a corpus of SMEI tweets using the fastText algorithm (Bojanowski, et al., 2017). The output of this model is a vector space, where each word has a semantic vectorial representation. The underlying idea is that encoded words "closer" in the vector space are expected to be similar in meaning. The dimension of this vector space is set to 300. Then, it is possible to map the input layers into a two-dimensional matrix: one dimension represents the word within the corpus and the other is its vectorial embeddings representation. This matrix is the input of the first LSTM layer and the subsequent output is the input of the second LSTM stage. The use of two stacked LSTM layers allows the model to capture the semantic relationships between words and sentences (Graves, et al., 2013). The first layer has 128 cells while the second 32. Both use a hyperbolic tangent (Tanh) activation function and a dropout rate of 0.5 for regularization. For dimensionality reduction, a 1-dimensional Max Pooling layer is then adopted to convert inputs (with various lengths) into a fixed-length vector. Finally, the output layer is a dense layer, i.e., a single fully-connected layer, which is a binary classifier. It uses a Sigmoid function, which is the predicted sentiment classification of each tweet: if the resulting quantity is higher than 0.5, then the tweet is classified as positive, otherwise negative.

The training process of the classifier (which allows the sentiment scoring of the output layers) can be carried out in a unique step or can be split into two phases, where the second step consists of a fine-tuning with the scope to specialize the classifier in a specific domain. To fine-tune the model, the dataset used for the one-step procedure needs to be split.

The fine tuning of the original model was carried out by using Italian economic tweets used for SMEI, while the retrained model for Covid uses the Feel-it dataset (Bianchi, et. al., 2021).

The original two-step model is pre-trained on a dataset composed by labelled Italian tweets coming from a variety of domains. The "pre-training" set is a merge of two datasets widely used for sentiment analysis, Sentipolc (Barbieri, et al., 2016) and Happy Parents (Mencarini, et al., 2019). The tweets within the training data include political and generic tweets, whereas the test data include tweets extracted with a socio-political topic via hashtags and keywords related to #labuonascuola. The Happy Parents is a dataset of Italian tweets related to parenthood. The merged dataset is composed of 6501 labelled tweets, the 39.44% are positive and the remaining are negative tweets. Then, the model is fine-tuned on a balanced set of 900 labelled tweets (year 2016) concerning economic topics used for internal uses in Istat. In the new retrained model, Istat dataset is always used in the first-step model.The fine-tuning of the model is performed by using an additional dataset: Feel-it consisting of 2037 tweets being the positive 35.73%. These tweets were retrieved by monitoring trending topics each day between 20th August to 12th October 2020, using the Twitter API. Feel-it dataset contains 662 COVID-19-related tweets.

Both datasets have been split into a training set and a validation set according to a proportion of 80/20. The model classification accuracy has been evaluated using the F1-Score. It is worth it to notice that these datasets are not recent and date back to 2016 in the best case. The trained model is then used to predict the sentiment of a set of 11,979,986 tweets in Italian referred to the period January - May 2020 extracted from Twitter by Istat by using a set of keywords related to economy as a filter. The predicted sentiment of each tweet is then used to build the daily index, which is computed as:

$$I = \frac{N_p - N_n}{N_p + N_n}$$

Where N_p is the share of tweets classified as positive each day while N_n is the share of tweets classified as negative. The same set of unlabelled tweets, together with other SMEI tweets othe months of April and May 2021 (for a total of 15.115.421 tweets), were used for theconstruction of the embedding Fastetxt space used as input layer.

3. Results

The original model achieves 0.80 as F1-Score on the validation dataset of the first step, 0.79 of the second step, while in the first step a 0.80. The retrained model achieves its highest accuracy on the validation of the fine-tuning set 0.86, 0.83 in the one-step and slightly less on first-step 0.79. When measuring the accuracy, with respect to the validation dataset, all models show high F1-Score values.

Table 1. F1-Score in the four scenarios.

Variable	With Feel-It	Without Feel-It
One-Step	0.83	0.8
Fine-Tuning (First step)	0.79	0.8
Fine-Tuning (Second Step)	**0.86**	0.79

The indexes created using the predicted sentiment in the different scenarios are illustrated in Figure 1, 2. The original model records a breakdown since the beginning of the Covid-19 pandemic, when a strong lockdown was imposed to the country. The index shows a downwards level-shift within the period between the 7th of March and the 21st of April, i.e., a full lockdown period in the country. In this time period it seems that one-step vs two-step procedure has the only effect of an upward translation. This is due to the fact in the fine tuning a balanced dataset (50% positive, i.e. mean 0) is utilized while in the one step procedure the dataset is unbalanced (40% positive, i.e. -0.2 index on average). The index, as shown in Figure 1, has some outliers, that need a deeper analysis. As expected, the maximum of the time-series is observed on the 1st of January. We analyzed the second maximum of the sentiment index on the 6th of March and the minimum on 11th of April. While the minimum value has a consistent meaning, the positive peak seems to be a false positive.

In the minimum value, s*pesa* (expense) is again among the most common words, a further confirmation that such term is correlated with negativity. The debate is focused around *mes* (the Italian word for the European Stability Mechanism), which appears to be negatively characterized in the twitter debate, as we observe other words such as *governo* (government), euro, *debito* (debt), *tasse* (taxes) in the conversations about MES.

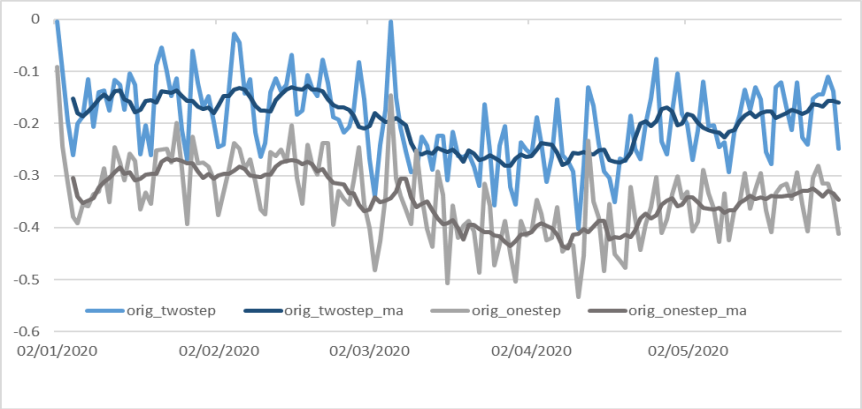


Figure 1: Original Model: daily index and weekly moving average (ma), in the two training processes (one step; two-step)

Concerning the positive peaks, we observe that when Coronavirus appears with the words *Italia*, *famiglia/e* (family), the tweets tend to be positively classified. The positivity linked to these words seems to confirm the intuition that the classifier assigns a sentiment according to the co-occurring words. Words as *famiglia/e* or *Italy* may have an intrinsic positive meaning, probably due the labelled data-set Happy Parents (more likely for family) or Sentipolc (for Italy). For this reason, we retrained the model.

In figure 2 we first observe that in this case the fine-tuning and the one-step procedure produce different results, and that the downward induced by Covid and Lock-down is more evident in the one-step trained model and is almost double than the original model or the two-step model. With respect to the original model both the re-trained models begin their descent since the 21st of February which is actually when the fear about the Italian spread of Covid began. With respect to the most positive values, we observe that in both cases, we have that second positive (or less negative) value is recorded on 13th April 2020 day on which Italy acceded to MES fund and deficit increase for economic restart was announced by the Government. It must be stresses that also the current SMEI index records such a positive value. While the minimum value in the two-step procedure is again the 11th of April in the one-step it is recorded on the 29th of March, day on which a new record of Covid deaths was witnessed. Finally, it must be stressed that the Feel-it dataset is the most negatively unbalanced dataset thus the two-step procedure lowers the average mean value of the index. Some useful additional insights can be obtained by analyzing the day over day variation of the weekly moving average (which is not a trend), as shown in Figure 3 and Figure 4.

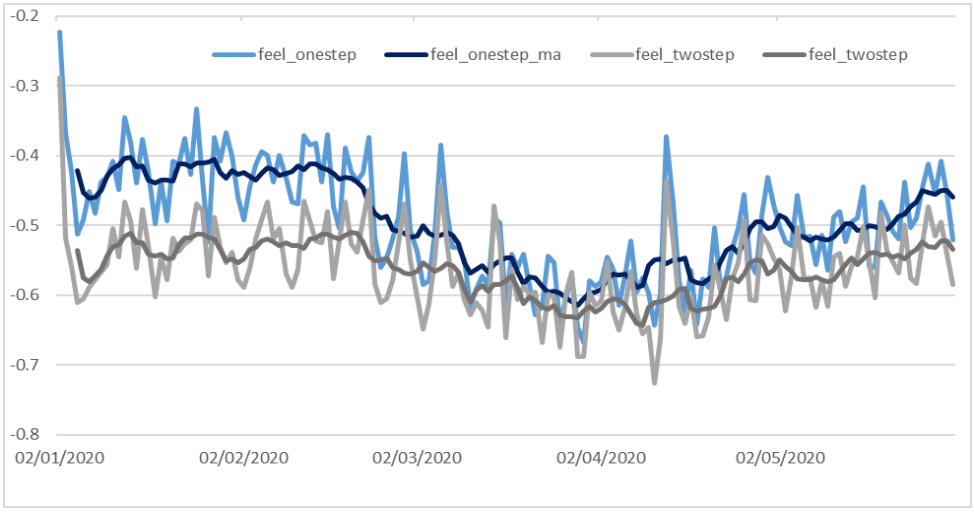


Figure 2: Retrained Model with Feel-it: daily index and weekly moving average, in the two training processes (one step; two-step)

In this case we observe that the retrained model with Feel-it shows since February 21st the same variations while they differ in the first part. In this sense the fine-tuning is able to specialize the model in Covid times as well as in the one-step procedure. When utilizing the whole dataset in the one-step procedure probably the rest of non-Covid related tweets (almost 1400) have the capability of changing the dinamics until late February.

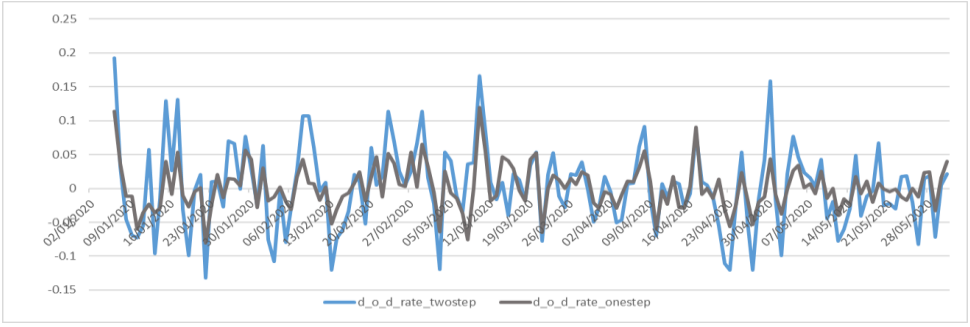


Figure 3- Day over day variation for original models.

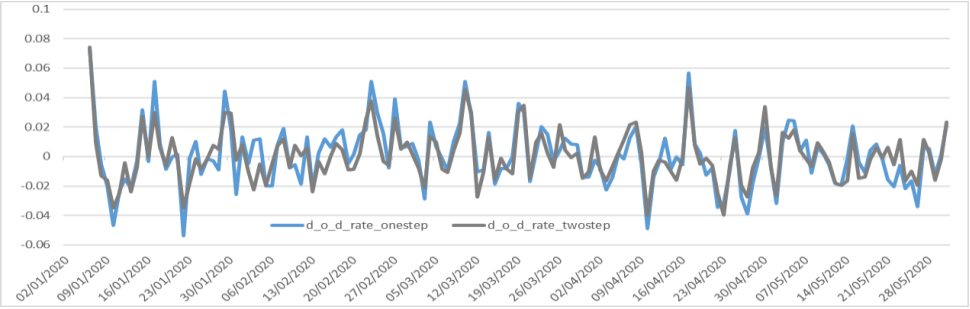


Figure 4- Day over day variation for Feel-it retrained models(lower)

If we analyse the original model, we observe that the one-step procedure smoothens the dynamics of the index as the day over day variations are always between the ones of the two-step.In the feel-it we observe an opposite behaviour.

4. Conclusions

The original index was already showing a breakdown during the Covid pandemic most likely induced by the fact that the input of the LSTM model is a word embedding model where Covid tweets have semantic relationships within SMEI tweets. However, we observed some misclassifications due to the training process. For this reason, we re-trained the model with a recent labelled dataset that contains lexical reference to the pandemic, e.g. Covid-19 terms, lockdown. Even if the size of Covid tweet is very modest we observe more coherently that the breakdown begins since late February 2020, and that the index decrease sharpens with respect to the original model as expected. It is not clear the specialization level provided by a two-step training. As a rule, it is advisable to increase the overall size of the annotated dataset in the first training step, so that fine-tuning could be performed to dynamically retrain the model on a relatively small dataset. In the present study we show that in case of Covid better results are obtained by a one-step procedure.

References

Barbieri, F., V. Basile, D. Croce, M. Nissim, N. Novielli, and V. Patti. 2016. “Overview of the Evalita 2016 SENTIment POLarity Classification Task”, *Proceedings of Third Italian Conference on Computational Linguistics*.

Bianchi, F., Nozza, D., & Hovy, D. 2021. “Feel-it: Emotion and sentiment classification for the italian language”, *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Bojanowski, P., E. Grave, A. Joulin, T. Mikolov. 2017. "Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics*, Volume 5: 135-146.

Catanese, E., M. Bruno, M. Scannapieco, and L. Valentino. 2022. "Natural language processing in official statistics: The social mood on economy index experience", *Statistical Journal of the IAOS*, Volume 38, Issue 4: 1-9.

Graves, A., N. Jaitly & A. R. Mohamed. 2013. "Hybrid speech recognition with deep bidirectional LSTM", *Proceedings of the 2013 IEEE workshop on automatic speech recognition and understanding*, pp. 273-278.

Hochreiter, S., & Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Huang, Z., Xu, W., & Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Medsker, L. R., & Jain, L. C. 2001. Recurrent neural networks. *Design and Applications*, *5*, 64-67.

Mencarini, L., Hernández-Farías, D. I., Lai, M., Patti, V., Sulis, E., & Vignoli, D. 2019. "Happy parents' tweets". *Demographic Research*, Volume 40, Article 25: 693-724.

Exploring emotional responses on Twitter after the Algeciras attack on Catholic churches in 2023: Between anti-immigration discourse and sadness reactions

Carolina Rebollo-Díaz[1], Estrella Gualda[1], Elena Ruiz-Ángel[1]
[1]Social Studies and Social Intervention Research Centre (ESEIS) / Centre for Research in Contemporary Thought and Innovation for Social Development (COIDESO), Department of Sociology, Social Work and Public Health, University of Huelva, Spain.

Abstract

On 25 February, a Muslim man attacked several churches in Algeciras (Spain) and killed a sexton. After the attack, many people turned to social media, especially Twitter, to express their emotions about what had happened, send their condolences to the deceased's family, or criticize the government, as the perpetrator was allegedly an undocumented migrant with a pending deportation order. The aim of this work is to study the emotional reactions of Twitter users who participated in conversations about the Algeciras case by applying sentiment analysis techniques. Using the academictwitteR package, more than 300,000 tweets containing the word 'Algeciras' were obtained. We then filtered out the RTs and kept 36,104 original tweets for this work. After data cleaning and tokenization, sentiment analysis was applied using the syuzhet package in R, which allowed to obtain the intensity of positive or negative sentiments and eight different emotions. The results suggest a higher prevalence of negative sentiments related to conversations about attacks, murder, or grief. The use of negative words reflects Twitter users' emotions, which are mainly concentrated on fear, anger, and sadness. Tweets expressing these emotions also indicated signs of Islamophobia and racism towards the murderer and, by extension, other Muslim immigrants.

Keywords: *Algeciras attack; anti-immigration emotions; sentiment analysis; syuzhet; Twitter; islamophobia.*

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16413

Account for variation by field in publication: bibliometric databases' analysis in a Portuguese Higher Education Institution

Cátia Malheiros[1], Conceição Gomes [1], Filipa Campos [1,2], Sofia Eurico[1]
[1]CiTUR, School of Tourism and Maritime Technology, Polytechnic of Leiria, Portugal. [2]CICF, Research Centre for Accounting and Taxation, School of Management, IPCA, Barcelos, Portugal

Abstract

This study seeks to examine the variation by field in publication practices in a Portuguese Higher Education Institution (HEI), where both research in Social Sciences and in Hard Sciences is conducted. The intention is to raise the issue of the suitability of bibliometrics for the Professors/researchers' evaluation considering their areas of research, as well as understanding the sort of use they make of these instruments. Different Bibliometric Databases were managed to analyze the use given to them by all the researchers in this HEI and to find out the main differences in its use according to the researched field of study. These results might represent a valuable source of information for HEIs in the process of finding the balance between the different procedures and format for the evaluation of researchers, to identify their in/ability to proficiently use these tools and to study the suitability of each tool to different profiles.

Keywords: *Evaluation; Higher Education; Bibliometrics; Social Sciences; Hard Sciences*

1. Introduction

In a recent past, bibliometrics have been assuming a crucial role in the evaluation process of Higher Education Institutions' professors/researchers, both in an individual basis as well as in a collective one, positioning the institution according to its information science results. However, the rapid evolution of bibliometric science and its close liaison to the evaluation of researchers does not make of the former expert users, or even interested ones in this method. Their evaluation was once a task led by peers and data are now "increasingly used to govern science" (Hicks et al., 2015:429) by Institutions, regardless the researchers' will or expertise in using these tools and the effectiveness of the service of bibliometric support research in libraries.

The concept of Bibliometrics has been used since 1969, when Pritchard (1969: 348) defined it as the "application of mathematics and statistical methods to books and other media of communication" and it dates back to the early 19th century, when the impact factor was firstly described by Eugene Garfield (1955) and when Tibor Braun launched the first dedicated journal Scientometrics in 1978 (Springer, 2023). Later on, and according to Furner, (2014:146), bibliometrics was described as being "about what people (authors, readers, etc.) do with documents (books, journal articles, web pages, tweets, etc.), for what reasons, and with what effects"

Introduction to Bibliometrics for the Evaluation of Scientific Information happens later on, on the threshold of the 21st century and reliance to its use is among much of the scientific community (Sugimoto & Larivière, 2018). The use of various research indicators should be done responsibly and ensure that these ones are not detrimental to the scientific community and that research is measured productively, supporting, rather than destroying, the scientific system (Sugimoto & Larivière, 2018). As Hicks et al. (2015:430) acknowledge and alert "Across the world, universities have become obsessed with their position in global rankings (…) even when such lists are based on what are, in our view, inaccurate data and arbitrary indicators". Moreover, the account for variation by field in publication is a concern that must be attended in order to avoid inequities and biases in the evaluation process (Nederhof, 2006).

According to the Leiden Manifesto for Research Metrics (Hicks et al., 2015) there are 10 ten principles to guide research evaluation, and the sixth one is directly linked with this topic. This study will therefore try to verify if the predominant area of study of the different researchers in a Portuguese HEI could be related to their presence and proficiency in the use of the metric databases at their disposal, leading to the following research question: What is the influence of the scientific field studied in the use of different bibliometric platforms by researchers to account for their publications?

2. Account for variation by field in publication

As clearly explained by Leiden (Hicks et al., 2015), quantitative metrics may not reflect with the same precision and justice the production of researchers from different areas of knowledge. Whether in the arts, social sciences or other areas, the studies that result from them have very specific nature and characteristics, which often do not match with publications that are plausible for bibliometric measurement. Working mainly on content, distancing themselves from quantitative instruments for the analysis of results, it is the social sciences that most resort to the use of qualitative methodologies, having often been kept away from publications in the 1st and 2nd quartiles and even recognizing a tendency to difficulty in being accepted in indexed journals for this very reason. According to College & James (2015:62), "the diverse nature of research at the institution as well as in the field should be highlighted, and appropriate denominators and indicators requested". This same idea is reinforced by Coombs & Peters (2017:8) about the Leiden manifesto, when saying that "the field normalization can be responsible for strongly influencing the result of the quantitative assessment, even more than the actual performance of the field".

The availability and willingness to use these platforms is often less among social science researchers who often publish their work in formats other than articles, which are the most easily measurable and accepted format for indexed publication. Some studies have been conducted in order to understand research output performance of social scientists as far as bibliometrics are concerned (Thanuskodi, S., 2017; Glänzel & Schoepflin,1999). For this study different bibliometric databases were considered with the purpose of gathering comprehensive research activity. Either researchers' unique digital identifier, as ORCID (Lehmann-Haupt, 2022), which allows the research to be guided by the individual, or bibliometrics databases, as Scopus (Scopus, 2023) which in turn guide research by the output results of researchers, have been used with the aim of enlarging the scope of this study.

3. Bibliometric databases' analysis in a Portuguese Higher Education Institution: methodological procedures

A Portuguese HEI which offers Bachelor's and Master's degrees in the field of Social Sciences (hereafter referred to as SSs) and that of Hard Sciences (hereafter referred to as HSs) and, therefore, having researchers from one area and the other equally, was chosen for the study. The Institution hosts two Research Units, one linked to SSs and one to HSs. The research units will be referred to as Group A (social sciences) and Group B (hard sciences).

All professors, working full time in this institution, teach and research simultaneously, and there is no separation of careers. From the 148 professors, only full members of the research units mentioned before and simultaneously working full time have been

considered, namely 35 (53%) in relation to the SSs one and 31 (47%) to the HSs one. The remain 15 professors are members of other research units not connected with the studied HEI.

Five different databases were explored during three months (December, 2022 to February 2023), namely: ORCID (ORCID, 2023), Scopus (Scopus, 2023), Web of Science (WOS) (Web of Science, 23), Dimensions (Dimensions, 2023), Google Scholar (Google Scholar, 2023) and 3 categories that were common in all the platforms: Articles in journals; Book chapters and Conference papers. Besides these 3 categories, some more were identified in ORCID and Google Scholar such as books, books edition, posters and patents. The information for each researcher was searched by name in all the mentioned databases and in ORCID. Sometimes, due to the difficulty in finding them by the name, it was necessary to add the institution or research centre to which they belong to.

Analysis and processing of data was achieved through excel and statistical package for social sciences (SPSS) software. Several descriptive statistics' measures were used such as median and variables were explored to analyse their normality and the existence of outliers through Kolmogorov-Smirnov test and Box-plot. As significance level of Kolmogorov-Smirnov test was <0.001, null hypothesis - the population is normally distributed - was rejected. Thus, as there is not normality, non-parametric tests were the option. Both research centres were compared based on publication practices using Mann-Whitney test (Pestana & Gageiro, 2014).

4. Results and discussion

4.1. Main differences in the use of database platforms according to the researched study field

By analysing the different database platforms, it is clear that the rate of publication in the SSs and HSs varies and Figure 1 shows the large discrepancy between the number of papers produced by researchers in Group A and those in Group B. It should be noted that ORCID and Google Scholar databases are those that bring together the largest number of publications by both groups. Considering that ORCID is the database that gathers the largest number of papers from both Group A and Group B, this base was selected to compare the two research groups in terms of quantity of publications.

Publication practices were compared between Group A and Group B researchers, using Mann-Whitney test. Null hypothesis (H0) was formulated: the distribution of total publications of *ORCID/Scopus/Web of Sciences/Google Scholar/Dimensions* is the same across group A and Group B. These hypothesis were rejected. Regarding these results, the difference between Group A and Group B is evident. The number of publications differs significantly between research centres, being necessary to determine which type of research

Group has the most publications in the analysed databases. Then, the median of the total publications was calculated in each Group for each database. The results show that the range of values of the median of publications of HSs is 41 to 22 and for SSs is 5 to 0. The difference between the groups was clarified, being the HSs researchers the ones with the highest number of publications.

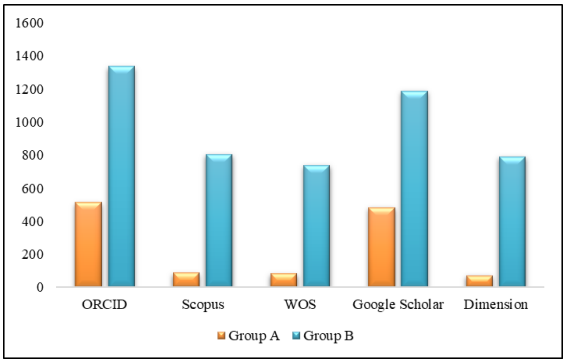


Figure 1 – Publications by database and research centre

While the normality was tested the variables were explored, and some outliers were found. This means that, in all the population, some researchers show up. They have so many publications that they become outliers, standing out from the overall pattern of the researchers considered for the study.

4.2. The variation by field regarding publication categories

Considering that ORCID is the most reliable database in terms of authorship, the typologies of documents in it were analysed in order to compare the differences in publication categories between research fields. Observing Table 1, in ORCID 72% of the documents were produced by Group B researchers and the predominance of publications in this group are articles in journals and posters. In the case of Group A, the typology with greater expression is the conference papers. Notice that there are neither patents in the case of Group A, nor edition of books in the case of Group B (Table 1).

ORCID database publication practices were compared between categories through Mann-Whitney test. Null hypothesis was formulated: the distribution of the articles in journals/books/book chapters/conference papers in ORCID is the same across the groups. For articles in journals and conferences papers this hypothesis was rejected. Moreover, the median of articles in journals highlights in Group B with the value of 26 against a value of 3 in Group A. In addition, the opposite happens in conferences papers, where Group A stands out with a median of 9 facing Group B which has a median of 2. With regard to books and book chapters H0 is not rejected, meaning that the distribution is similar between both groups.

Table 1. Publication categories in *ORCID*

ORCID	Group A	Group B	Total	Group A	Group B
Articles in journals	165	797	962	17%	83%
Books	28	40	68	3%	4%
Book chapters	67	78	145	7%	8%
Posters	3	351	354	0%	36%
Conference papers	240	54	294	25%	6%
Book's edition	10	3	13	1%	0%
Patents	0	12	12	0%	1%
Total	513	1335	1848	28%	72%

Furthermore, HSs researchers record the highest number of articles in journals. However, SSs researchers have the highest number of conference papers. These researchers have different profiles according to the field that they belong to, being undeniable the higher number of the publications of HSs' researchers.

5. Conclusions, limitations and further research

From the obtained results, it is clear that HSs' researchers present more publications than those from the SSs and that, in both areas, some publish more than others within the same group. There are also differences in the typology of documents produced by each group. Journal articles are very high in Group B and they also publish many posters and have patents. On the other hand, conference papers are higher in group A. These results corroborate the research question of the study, that inquires the influence of the scientific area studied and its relationship with publication performance in different bibliometric platforms.

As for the studied platforms, ORCID presents itself as the one with the larger number and diversity of documents and both ORCID and Google Scholar gather the greatest diversity of documents. Scopus, WOS and Dimension only consider three types of documents, disregarding other works for the indicators.

A careful reflection of the obtained results, and in the light of the theoretical framework, highlights that a proper and legitimate diversity of the different scientific areas requires a correct and dignified treatment of scientific production, regardless of its nature or format. Scientific production demands, in its essence, quality and accuracy and these should prevail over formatting standards, style and methodologies that are imposed and that, ultimately, do

not dignify the research product *per si*. The scientific production of the SSs ends up having less expression and visibility in the platforms. Their lower presence may not be an indicator of lower productivity, but perhaps of the lack of systematisation and registration of production in these instruments or the inadequacy of the parameters and formats required equally for all areas, not looking at their specificities.

Moreover, the danger of the commercialising of science by imposing practices for measuring results that do not always match the nobility, breadth and diversity of types of studies of scientific production may result from an incorrect use of these instruments. If they are not seen as an auxiliary measuring mechanism, instead of a prevailing instrument to validate scientific production, we may move towards a pathogenic culture that, according to Mendon (1942), results from an imperative logic of publication as a way of belonging to the community, and that is frequently cause for fraudulent behaviour.

As for the limitations of the present study, in fact they somehow enhance future research. For instance, in some cases, difficulty in identifying the researcher in the different databases, due to the absence of a profile or publication or even the presence of more than one profile for the some researcher, leads to the need of a future study that compares different databases and author identifiers and recognizes weaknesses and advantages among them.

This study compares 4 databases and 1 author identifier, for a singular HEI and the development of similar studies, but in a broader context, including different Portuguese HEIs and even expanding it to a worldwide context, would be advisable. Further research should also consider the implementation of new instruments for database assessment and better performance of the analysis, through more advanced artificial intelligence.

Funding: This work is financed by national funds through FCT - Foundation for Science and Technology, IP, within the scope of the reference project UIDB/04470/2020.

References

Colledge, L. & James, C. (2015). “A “basket of metrics”—the best support for understanding journal merit”, *European Science Editing*, *41*(3), 61-65.

Coombs, S. K. & Peters, I. (2017). The Leiden Manifesto Under Review: What Libraries Can Learn From It. *Digital Library Perspectives*. https://libereurope.eu/wp-content/uploads/2020/11/DLP-Paper.pdf

Dimensions. (2023). *Linked research data from idea to impact*. Digital Science & Research Solutions Inc. https://www.dimensions.ai/

Furner, J. (2014). The Ethics of Evaluative Bibliometrics. In CRONIN, Blaise; SUGIMOTO, Cassidy R., eds. (2014). Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. Cambridge: MIT, 85-107.

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, *122*(3159), 108-111. http://www.garfield.library.upenn.edu/papers/science_v122v3159p108y1955.html

Glaènzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing and Management, 35*(1), 31-44. https://www.sciencedirect.com/science/article/pii/S0306457398000284

Google Scholar (2023). Google Scholar. https://scholar.google.com/

Hicks, D. & Wouters, P. (2015). The Leiden Manifesto for research metrics. *Nature*, 520, 429-431.

Lehmann-Haupt, J. (2022). *ORCID´s Frist decade: From Startup to Sustainability*. ORCID. https://info.orcid.org/wp-content/uploads/2022/11/R2_Orcid-10th-Ann-Booklet-FOR_WEB.pdf

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities, *Scientometrics*, 66(1), 81-100.

ORCID. (2023). *ORCID – Connecting research and researchers*. https://orcid.org/

Pestana, M. & Gageiro, J. (2014), Análise de Dados Para Ciências Sociais, 6th ed., Sílabo.

Pritchard, A. (1969). Statistical Bibliography or Bibliometrics. *Journal of Documentation, 25*(4), 348-349.

Scopus (2023). *Start exploring - Discover the most reliable, relevant, up-to-date research*. Elsevier B.V. https://www.scopus.com/search/form.uri?display=basic#basic

Springer (2023). *Scientometrics - An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy*. Springer Nature Switzerland AG. https://www.springer.com/journal/11192/editors.

Sugimoto, C.R. & Larivière, V. (2018). *Measuring research: what everyone needs to know*. New York: Oxford University Press.

Thanuskodi, S. (2017). Journal of Social Sciences: A Bibliometric Study. *Journal of Social Sciences, 24*(2), 77-80. https://www.tandfonline.com/doi/abs/10.1080/09718923.2010.11892847

Web of Science. (2023). *Web of Science Platform - Training Resources*. Clarivate Accelerating innovation. https://clarivate.com/webofsciencegroup/support/wos/

Emotions of the main educational agents involved in the App educational applications

Francisco Javier Rondán-Cataluña[1], Begoña Peral-Peral[1], Patricio E. Ramírez-Correa[2], Jorge Arenas-Gaitán[1]
[1]Department Business Management and Marketing, University of Seville, Spain, [2]Catholic University of the North, Chile.

Abstract
The integration of digitization into various aspects of daily life has been accelerated recently, particularly in the realm of e-government. This study focuses on examining the emotions of key stakeholders in non-university public education, specifically educational centers, teachers, and families, as they pertain to educational applications developed by Spanish autonomous communities. The research employs a novel approach, incorporating word processing analysis to evaluate the emotions expressed in Twitter posts by the aforementioned groups. The analysis employs the Plutchick model of emotions and feelings, utilizing various R libraries designed for this type of analysis. The findings suggest differing perceptions of educational apps among the studied groups.

Keywords: *analysis of emotions, sentiment analysis, Twitter*.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16456

Newspapers, images and income support policy

Pietro Cruciata[1,2], Chiara Perfetto[2], Giuliano Resce[2]
[1]Polytechnique Montréal, Canada [2] Department of Economics, University of Molise, Italy

Abstract

To what extent do different newspapers have different kinds of images associated with articles on the same topic? We investigate this research question by considering one of the most important Income Support Policies implemented in Italy in recent times ('Reddito di cittadinanza' – RdC) which generated a strong debate in public opinion. Focusing on the national wide media, we downloaded images associated with articles about RdC and by means of Image Captioning algorithms, we generate the description of them. Results show that different newspapers have images containing different objects. Some topics emerging from images published by newspapers are very exclusive and the sentiment associated with the text extracted from the images has a wide heterogeneity. Furthermore, right-hand newspapers show a lower sentiment compared with left-hand newspapers. Overall, the results confirm that the ideological stance associated with different media outlets is reflected also in the images associated with articles and that the integration of Image Captioning algorithms and Natural Language Processes is very promising in this research area.

Keywords: *Image Analysis; Text Mining; Political Debate; Income Support.*

1. Introduction

It has been widely shown that media outlets have their own ideological stance, which implies a bias in the spreading of news, such as how the topic is covered and how it is presented and discussed (Le Moglie, Turati, 2019; Gentzkow, Shapiro, 2010, Mullainathan, Shleifer, 2005). The heterogeneity in the media's ideological stance has a crucial role in the quality of a democracy. Still, it has also been shown that, since profits are driven by the number of readers or viewers, the supply and the discussion of some news could be caused by users' preferences, leading to the reduction in the supply of information more relevant for the accountability of the political system and less attractive for the public (Ho, Liu, 2015, Sen, Yildirim, 2015).

In this paper, we are interested in understanding to what extent different newspapers have different kinds of images associated with articles on the same topic. We investigate this research question by considering a topic that generated a strong debate in Italian public opinion: an Income Support Policy called '*Reddito di cittadinanza*' (RdC). We build a new original database containing all the images associated with articles about '*Reddito di cittadinanza*' published by all the Italian national-wide newspapers. There is a specific reason why we expect heterogeneity in the images associated with the articles: the debate on the RdC has been characterized by a high degree of politicization. In the 2022 Italian electoral campaign, the RdC was one of the main characters with the programs of the various parties proposing different actions on the policy. The Five Stars Movement (M5S) wanted to strengthen the current system, the Democratic Party (PD) and the Third Pole propose relevant reforms while the unitary program of the Centre-right plans propose to replace the RdC with alternative measures of social inclusion.

Preliminary results show that the ideological stance associated with different media outlets is reflected also in the images associated with articles and that the integration of Image Captioning algorithms and Natural Language Processes is very promising in these analyses. Different newspapers have images containing different objects, and the sentiment associated with the text extracted from the images has a wide heterogeneity, in particular, right-hand newspapers show a lower sentiment compared with left-hand newspapers.

2. Data and method

2.1. Data

To create the corpus for analysis, we developed a Python web scraper. Our objective was to search and download all images related to the query "reddito di cittadinanza" from Google, while targeting specific newspapers adding to the query the name between quotation marks to get more precise results. We decided to gather images from Google for a maximum of 300

images per newspaper. We limited the collection to a maximum of 300 images per newspaper, as we observed that beyond this threshold, the relevance to our research decreased. From Google, we retrieved both the images and the website that uploaded them, and we ensured that an image is associated with a unique link to the respective article. Then, we filtered out any images that did not originate from the selected newspapers' websites. Finally, our sample includes 474 images from 9 different newspapers and are distributed as follows in Table 1:

Table 1. Image distribution for each newspaper

Newspaper	Number of images
Il Mattino	66
La Repubblica	91
La Stampa	20
Il fatto quotidiano	90
Il Corriere della Sera	79
Il Messaggero	71
Il Sole 24 Ore	31
Libero	20
Tuttosport	6

2.2. Methodology

Image Captioning aims to briefly describe an image to assign it a caption. We use an algorithm developed in this subfield extending the captions generated to have a more complete description of the images studied.

Researchers created Image Captioning algorithms through the combination of state-of-the-art algorithms in two main fields of AI: Computer Vision and Natural Language Processing (NLP). Computer vision algorithms are used to recognize the entities in an image while NLP algorithms are used to generate the description of it.

The algorithm that we propose is the Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BLIP) (Li, Li, Xiong, & Hoi, 2022). This model is based on two pre-trained transformers: the image transformer initialized from ViT pre-trained on ImageNet (Dosovitskiy et al., 2020), and the text transformer initialized from BERT base (Devlin, Chang, Lee, & Toutanova, 2018). Moreover, at the core of BLIP there is a multi-task pre-training and flexible transfer learning architecture called Multimodal

mixture of Encoder-Decoder (MED). MED works pre-training at the same time three different vision-language objectives: image-text contrastive learning, image-text matching, and image-conditioned language modeling. Finally, MED is finetuned with Captioning and Filtering: a new dataset bootstrapping method in which a captioner produces synthetic captions given web images, and a filter removes noisy captions from both the original web texts and the synthetic texts.

The model provides the flexibility to set various parameters, including the number of beams used for the beam search, the option to use nucleus sampling, and the minimum and the maximum number of characters generated for each caption. Nucleus sampling and beam search are two types of algorithms that allow the model to generate the caption words. While the beam search generates the most probable sequence of tokens, nucleus sampling has a different approach. It generates a subset of tokens where the sum of their probabilities is greater than a predetermined value. For this pilot study, we generate 20 captions for each image, one using beam search and 19 using the nucleus sampling. Additionally, we set the minimum number of words to 10 and the maximum to 20.

Thus, we proceed with the image captions analysis, which will involve unsupervised text mining and statistical analysis of captions to extract information, as well as sentiment analysis to detect the semantic orientation of the content. The different captions of each image will be aggregate/paste into one big text and will be represented by the term document matrix (TDM). This representation allows the data to be analyzed with vector and matrix algebra, effectively moving from text to numbers. In the TDM the rows correspond to the terms in the caption, columns correspond to the newspapers and cells correspond to the frequency of the terms. Not all terms are equally informative for text analysis (Welbers et al., 2017). One of the first things to remove very common terms is the use of "stopword" lists, but it is not sufficient there may be still other common words, and this will be different between corpora. For that, an additional approach is to assign them variable weights. A popular weighting scheme is the term frequency-inverse document frequency (TF-IDF), which uses information about the distribution of terms in the corpus to estimate how exclusive the association is between a word and a document (Hidalgo and Hausmann, 2008; 2009). In detail, TF measures the term-frequency that is the times that a term occurs in the given document and IDF indicates how common and rare that term is across all documents. Formally:

$$IDF(term) = ln\left(\frac{n_{documents}}{n_{documents\ containing\ term}}\right) \quad (1)$$

A high TF-IDF value indicates that the term is important to the given document and possibly represents key information o that document.

For sentiment analysis of the text coming from the captions, we will use the "syuzhet" package (Jockers, 2017) in R which allows the calculations of polarity scores of a collection

of documents by using different internal dictionaries. We will focus on the "nrc" dictionary developed by Turney and Mohammad (2010). The words in the dictionary are classified into multiple labels corresponding to positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust. The algorithm after the comparison between the words in the text with the words in the dictionary counts them and returns a data frame in which each row represents the captions. The columns include one for each emotion type as well as the positive or negative sentiment valence. The score obtained in each cell is divided by the total number of words in each caption. In this way, each image will correspond to a text and the text will associate with emotions.

3. Preliminary results

3.1. Discussion

Figure 1 reflects the TF-IDF of the terms characterizing each newspaper. The right-hand newspaper like "Libero" focuses on "wallet" and "euro"; "La Repubblica" e "Il Mattino" are the most similar and both focuse on "card" and "credit". "La Stampa" and "Il corriere della sera" report image of the police, while "Il fatto quotidiano" rappresents the "people". The image from "Il Sole 24 Ore," which is an economic-centric journal, appears to be depicting images that are more closely related to politicians. However, these images may not be informative in our analysis, as the model often struggles to recognize the individuals depicted. Similarly, "Tuttosport," a sports newspaper, is characterized primarily by sports-related images, which are not relevant to our research. Therefore, these last two newspapers will not be included in the subsequent sentiment analysis as they are unrelated to the RDC issue as demonstrated by TF-IDF.

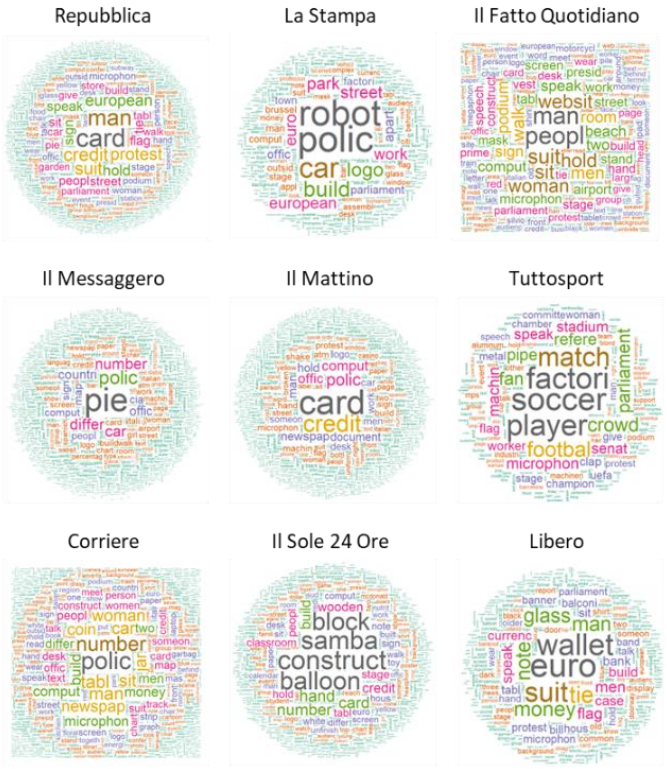


Figure 1. TF-IDF by newspaper

The results obtained from the sentiment analysis are shown in Figure 2 and in Figure 3. Figure 2 shows the percentage of words associated with eight different emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) plus negative and positive general sentiments. The emotions are distributed on the x-axis from the most negative (anger) to the most positive (trust). It is evident from Figure 2 that the emotion "trust" has the longest bar indicating that words associated with this emotion constitute about 20% of all the meaningful words in this text. On the other hand, the emotion of "disgust" shows that words associated with this negative emotion represent about 2% of all the meaningful words in this text. In general, the sentiment is more positive than negative. The specific sentiment related to each newspaper is shown in Figure 3.

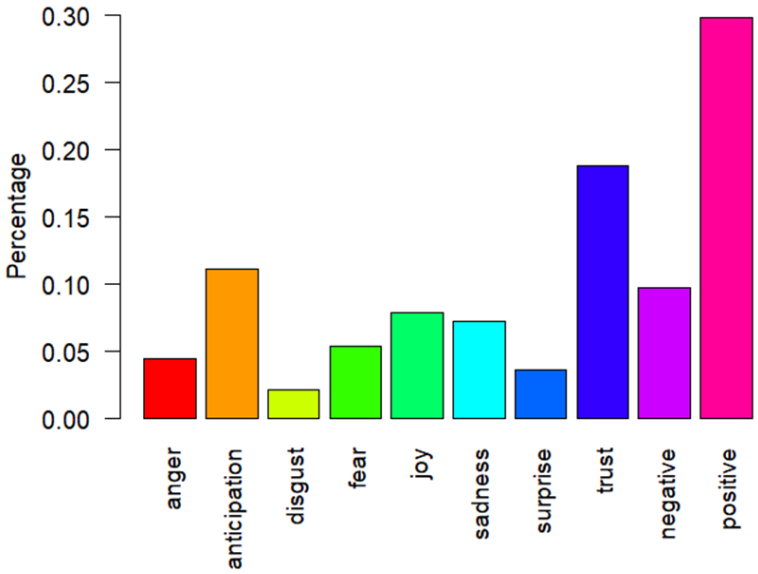


Figure 2. Percentage of words in the text associated with each emotion

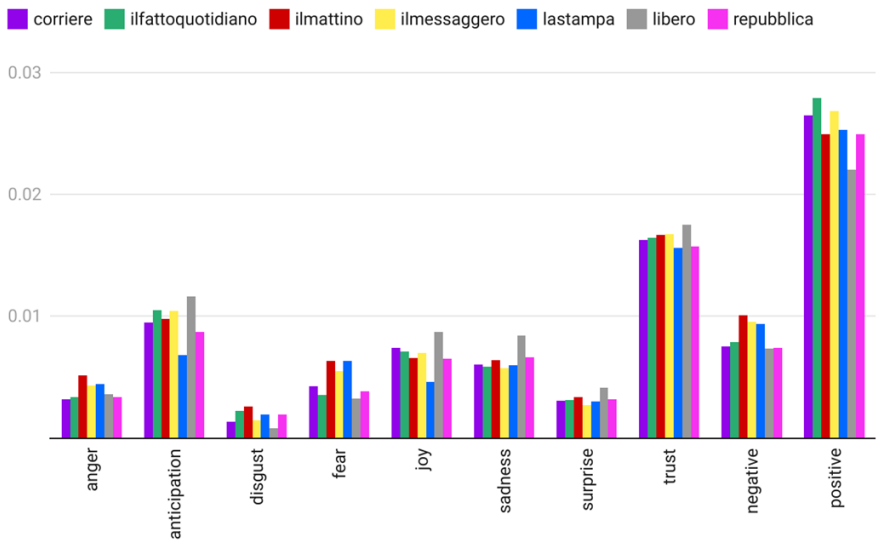


Figure 3. Presence of emotions in each newspaper

Figure 4 shows how the sentiment of each newspaper differs from the average sentiment. The sentiment relating to "Il Fatto Quotidiano" differs positively from the general sentiment average. The lowest sentiments are associated with images in "Il Mattino" and "Libero".

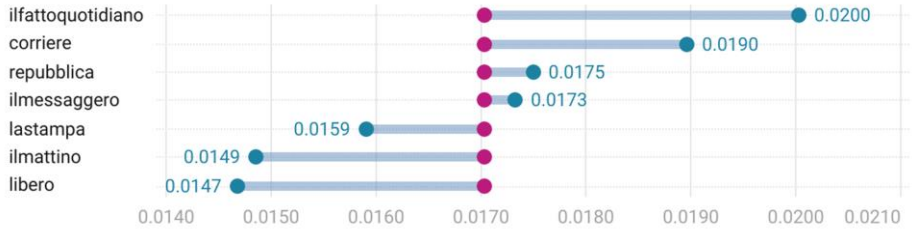


Figure 4. Deviation of the sentiment associated with each newspaper from the average sentiment

In summary, the results confirm that the ideological stance associated with different media outlets is reflected also in the images associated with articles, and in particular the right-hand newspapers show a lower sentiment compared with left-hand newspapers.

3.2. Limitation and Future research

Despite the good results, there are some limitations to this work that surely affect our results. The first limitation is not the best quality of the images. This surely prevents sometimes the ability of the BLIP model to spot all the objects in an image to fully describe it. Another limitation this time is on the language generation part. The model comes already with pre-acquired knowledge that is built in English. Thus, it has difficulties to interpret the image and the context of the topic that we are trying to analyze. Another bias in our methodology stems from the missing data in the information retrieved. Indeed, the images described could bring interesting information to the topic study.

Building from these limitations, there are two main improvements possible to improve the results of our work. The first is to create a pre-trained model that is trained on the subject. This will allow the model to understand the story and the subject involved easing the interpretations. The second improvement possible is to add the date of the image published online to analyze the event that is connected to the different images published by the examined journals.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Preprint ArXiv:2010.11929*.

Feinerer I, and Hornik K (2018). tm: Text Mining Package. R package version 0.7-5.

Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica*, 78(1), 35-71.

Hidalgo, CA, and Hausmann, R (2008). A network view of economic development. *Developing alternatives*, 12(1), 5-10.

Hidalgo, CA, and Hausmann R. (2009). The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26), 10570-10575.

Ho, B. and P. Liu (2015). Herd journalism: Investment in novelty and popularity in markets for news. *Information Economics and Policy* 31, 33–46

Jockers, M. (2017). syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text. R package version 1.0.6.

Le Moglie, M., & Turati, G. (2019). Electoral cycle bias in the media coverage of corruption news. *Journal of Economic Behavior & Organization*, 163, 140-157.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ArXiv Preprint ArXiv:2201.12086*.

Mullainathan, S., & Shleifer, A. (2005). The market for news. *American economic review*, 95(4), 1031-1053.

Mohammad, S., & Turney, P. (2010, June). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. *In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26-34).

Sen, A. and P. Yildirim (2015). Clicks and editorial decisions: How does popularity shape online news coverage? *Mimeo*. Available at SSRN

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication methods and measures*, 11(4), 245-265.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16433

Food insecurity trends in the Famine Early Warning Systems Network

Bia Carneiro[1], Chiara Perfetto[2], Giuliano Resce[2], Giosuè Ruscica[1], Giulia Tucci[1]
[1]Alliance of Bioversity International and CIAT, [2]Department of Economics, University of Molise, Italy.

Abstract

Over last 30 years, periodic country analyses elaborated by FEWS NET (Famine Early Warning Systems Network of the United States Agency for International Development) enabled creation of a unique source of knowledge comprising consistent reporting in over two dozen countries. This paper proposes to systematically assess documentation from historical perspective to provide comprehensive overview of food insecurity in FEWS NET covered countries. We propose an integrated machine learning approach to systematically analyse available documentation and generate knowledge. In particular text mining algorithms have been implemented to analyse reports: automated retrieval of high-quality information from text, by finding patterns and trends through machine learning, statistics and linguistics. This enables analysis of large amounts of unstructured text to derive insights. Results show that there is a wide heterogeneity in what is relevant, and in what reports focus on at the territorial level. Many country-level topics are persistent over time with some interesting exception, as Guatemala, Malawi, Niger, and Somalia with more instability. Overall, the evidence show that advances in machine learning and Big Data research offer great potential for international development agencies to leverage the vast information generated from reports to gain new insights, providing analytics that can improve decision-making.

Keywords: *Food insecurity; Early Warning Systems; Text Mining.*

1. Introduction

The Famine Early Warning Systems Network (FEWS NET) is a leading provider of early warning and analysis on acute food insecurity around the world. Created in 1985 by the United States Agency for International Development (USAID) in response to devastating famines in East and West Africa, FEWS NET provides evidence-based analysis to governments and relief agencies who plan for and respond to humanitarian crises. FEWS NET analyses support resilience and development programming as well. FEWS NET analysts and specialists work with scientists, government ministries, international agencies, and NGOs to track and publicly report on conditions in the world's most food-insecure countries.

The FEWS NET reporting includes:

- Monthly reports and maps detailing current and projected food insecurity
- Alerts on emerging or likely crises
- Special reports on factors that contribute to or mitigate food insecurity, including weather and climate, markets and trade, agricultural production, conflict, livelihoods, nutrition, and humanitarian assistance
- Access to data, learning, and analysis of the underlying dynamics of recurrent and chronic food insecurity and poor nutritional outcomes, to improve early warning and better inform response and program design

Over last 30 years, periodic country analyses elaborated by FEWS NET enabled creation of a unique source of knowledge comprising consistent reporting in over two dozen countries (Figure 1). Longitudinal breadth of knowledge has potential to provide information and insights into long-term trends regarding shocks over time, coping capacity, and livelihoods. However, time and resource constraints and focus on current humanitarian assistance mean such outputs underutilized as sources of evidence.

This paper proposes to systematically assess documentation from historical perspective to provide comprehensive overview of food insecurity in FEWS NET covered countries. The insights can also support current operations by providing another evidence-base from which to carry out analyses.

Extracting knowledge from extensive text-based sources poses several challenges (Garbero *et al.* 2021). Conventional approaches such as manual coding or keyword searches time and resource intensive (Carneiro *et al.* 2022). This is the reason why we propose a text mining approach: automated retrieval of high-quality information from text, by finding patterns and trends through machine learning, statistics and linguistics (Resce, Maynard, 2018). This enables analysis of large amounts of unstructured text to derive insights. Our proposal is an

integrated machine learning approach to systematically analyse available documentation and generate knowledge.

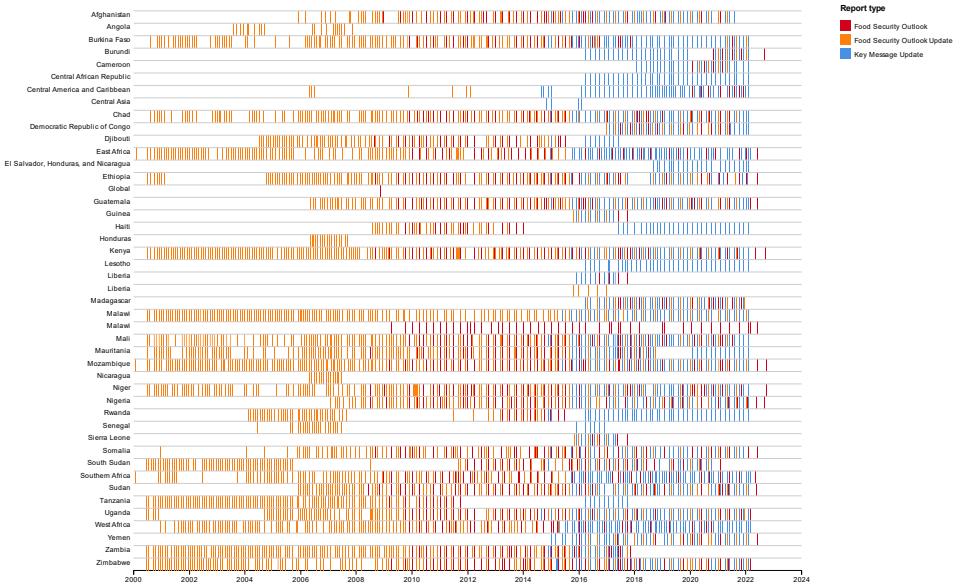


Figure 1 FEWS NET Reports Timeline

The research questions of the present analysis are the following:

- What historical trends (or anomalies) can be identified regarding the prevalence of various dimensions from the FEWS NET framework for acute food insecurity (i.e. reported hazards/shocks and outcome/impacts, particularly focused on indicators related to infectious diseases, agroclimatology, markets, and conflict)?
- How do trends compare by region and type of hazard/shock?
- How are the dynamics, such as the interactions within and between the different dimensions, represented? How do these representations transform in time and space?

2. Material and methods

We start systematizing the 5217 total Documents available in FEWS NET (1017 Food Security Outlook; 2923 Food Security Outlook Update; and 1276). For each report in machine readable PDF format, relevant sections were extracted into spreadsheet with section headers and column title. We only focus on the section National overview, which contains the description of the present situation of the country the report is referring to.

The corpus of analysis was prepared using functions from the R package "tm" (Feinerer, Hornik, 2018; Feinerer *et al.* 2008): punctuation, stop words (i.e. in English, words like "the", "is", "of", etc), and numbers were removed from the corpus. The words were then converted to lowercase and stemmed. The most common formats for representing a corpus of texts (i.e., a collection of texts) in a bag-of-words format is Term Document Matrix (TDM). A TDM is a matrix in which rows are terms, columns are document/report, and cells indicate how often each term has occurred in each document/report. The advantage of this representation is that it allows you to analyse data with vector and matrix algebra, effectively moving from text to numbers. In the TDM the frequently occurring terms are assigned a higher score than the rarely occurring terms. Starting from TDF with the number of occurrences it is possible to quantify what a document is about. The TF-IDF score is another useful metric used to populate the TDM. One measure of a word's importance is its term frequency (TF), which counts a word's occurrence in a document. Another approach is to look at a term's inverse document frequency (IDF), which decreases the weight of commonly used words and increases the weight of words that do not appear frequently in a collection of documents. The two can be combined to calculate a term's TF-IDF (the two quantities multiplied together), which measures the frequency of a term adjusted for how rarely it is used (Silge, Robinson, 2017). Formally:

$$IDF(term) = ln\left(\frac{n_{documents}}{n_{documents\ containing\ term}}\right) \quad (1)$$

In our case, the TF-IDF combines frequency, i.e., how many times a word is associated to a document, and the inverse of ubiquity, i.e., how exclusive the association is between a word and a document (Hidalgo and Hausmann, 2008; 2009). To this regard, it is worth stressing that more ubiquitous words are more likely to have less informative power than exclusive words.

From the TDM perform correlation analysis to understand to what extent does the usage of a topic change (increase or decrease) in relation to the usage of the same topic in a previous text. A correlation analysis aims to determine the extent to which there is a relationship, or linear dependence, between two sets of points (Jockers, 2014). The correlation is a measure of the strength of the linear dependence between the word frequency in one report and the word star frequency in another report. This result, called the Pearson moment-product absorbing coefficient, is expressed as a number between -1 and +1. A negative coefficient one (-1) negative represents a perfect negative; if the correlation between the reports is -1, then we would know that the higher frequency of a word in one report corresponds proportionally to the lower frequency of the same word in the other report. This means that something changes in the report A positive one (+1) represents a perfect positive correlation (when one variable goes up and down, the other variable does so ideally).

3. Results and Conclusions

Figures 2, 3, 4, and 5 show the TF-IDF word-cloud by countries, the shape of the clouds reflects the map of the country. The figures are organised by Regions: Figure 2 is East Africa, Figure 3 is Southern Africa, Figure 4 is West Africa, and Figure 5 is MENA & LAM.

Figure 2 TF-IDF by country (East Africa)

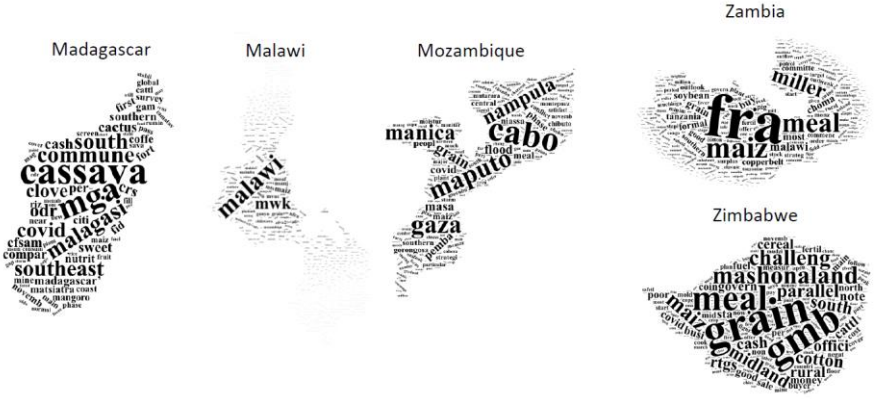


Figure 3 TF-IDF by country (Southern Africa)

Figure 3 TF-IDF by country (West Africa)

Figure 4 TF-IDF by country (MENA & LAM)

Overall, Figures 2-5 show that there are specific terms associated to each country, meaning that there is a wide heterogeneity in what is relevant, and in what reports focus on at the territorial level. Figure 6 shows the average year correlation for a subset of countries having reports in almost all years. The correlation is estimated between a report and the previous report for the same country. We report year average to clean up the analysis for the seasonality that is intrinsic in many events. Figure 6 shows that the correlation is quite high in almost every country, meaning that country-level topics are persistent over time, but there are some interesting exception, such as Guatemala, which shows a flat low correlation. Other countries show a non linear trend, suggesting an internal instability, they are: Malawi, Niger, and Somalia.

Overall, the preliminary evidence show that advances in machine learning and Big Data research offer great potential for international development agencies to leverage the vast information generated from reports to gain new insights, providing analytics that can improve decision-making.

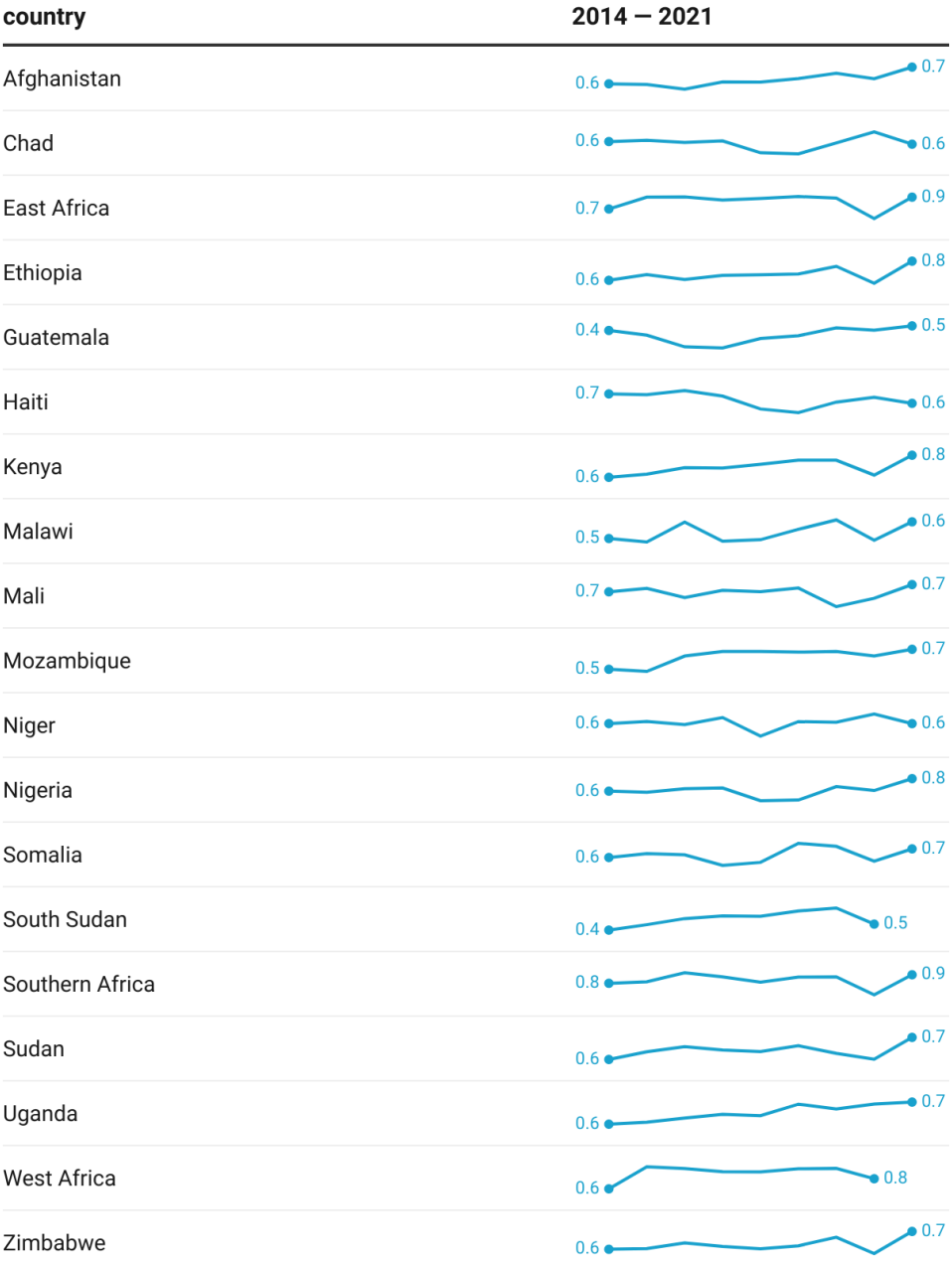


Figure 5 Annual trend of correlation with previous report for a subset of representative Country

References

Carneiro, B., Resce, G., & Sapkota, T. B. (2022). Digital artifacts reveal development and diffusion of climate research. *Scientific Reports*, 12(1), 14146.

Feinerer I, and Hornik K (2018). tm: Text Mining Package. R package version 0.7-5.

Feinerer I, Hornik K, and Meyer D (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5): 1-54. URL: http://www.jstatsoft.org/v25/i05/

Garbero, A., Carneiro, B., & Resce, G. (2021). Harnessing the power of machine learning analytics to understand food systems dynamics across development projects. *Technological Forecasting and Social Change*, 172, 121012.

Hidalgo, CA, and Hausmann, R (2008). A network view of economic development. *Developing alternatives*, 12(1), 5-10.

Hidalgo, CA, and Hausmann R. (2009). The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26), 10570-10575.

Jockers, M. L. (2014). *Text Analysis with R for Students of Literature*. Cham: Springer International Publishing.

Resce, G., & Maynard, D. (2018). What matters most to people around the world? Retrieving Better Life Index priorities on Twitter. *Technological Forecasting and Social Change*, 137, 61-75.

Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. O'Reilly Media, Inc.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16463

0-shot text classification for web-based environmental indicators: Pilot study on B-Corp data

Pietro Cruciata[1], Davide Pulizzotto[1], Mikaël Héroux-Vaillancourt[1], Catherine Beaudry[1]
[1]Polytechnique Montréal, Canada

Abstract

This paper proposes a tool that uses web-based information to generate a proxy for the environmental culture indicator developed by B-Lab. The tool is based on recent advances in Natural Language Processing (NLP), such as pre-trained language models like BART that better capture the semantic facets of natural language. The algorithm and data provide several advantages, including real-time analysis, minimal building cost, granularity, and a large sample size, making it appealing. The Zero-shot text classification task is used to create an indicator of companies' environmental culture, which was chosen due to the urgency created by recent climatic events, pushing for increased environmental protection and sustainability culture promotion. The tool was tested on the B-CORP dataset, which provides scores on environmental performance. Results indicate that scores for certain environmental topics generated by the tool are correlated with B-Lab's environmental indicator. This research open door to the possibility of predicting the environmental readiness of the companies base on web-based indicators.

Keywords: *Natural Language Processing, Zero-shot text classification, Sustainable Innovation*

1. Introduction

Governments acknowledge that innovation is a critical driver of economic growth, and thus, they allocate funds to support companies' research and development (R&D) projects. If the government can allocate these funds efficiently, it can lead to accelerated economic development. Typically, policy makers and governments rely on administrative data and questionnaire-based surveys to assess the quantitative impact of these R&D investments. While these sources of information may serve their intended purpose, they often have significant limitations. For instance, connecting particular policy instruments to alterations in firm performance, as assessed by administrative data, poses a challenge. Furthermore, surveys reliant on questionnaires (particularly those on a large-scale, such as the biennial European CIS or the annual MIP) are inadequate in terms of regional granularity, scope, timeliness, and conducting such surveys incurs significant costs. (Axenbeck and Breithaupt 2021). Due to all these factors, conventional indicators of innovation seldom offer a comprehensive view of the effects of policy combinations (Kinne and Lenz 2021). Alternative or complementary to these sources are web-based unstructured textual data. Among their advantages, the rapidity of their evolution, their increasing quantity, variety, and availability opened new possibilities for policy makers and researchers (Gök, Waterworth, and Shapira 2015). As policy makers now turn their attention to adaptation to climate change, mitigation of its effect, and generally a better socio-environmental impact of their policies, sustainable innovation is perceived as a key solution.

This paper proposes a method that employs web-based information to create an environmental culture indicator proxy of the real environmental culture indicator developed by B-Lab. Indeed, the environmental impact of the companies plays a crucial role in the triple bottom line framework established by John Elkington in the 1990s, which highlight the equal importance of social, environment and economics goals to pursue sustainable innovation. Moreover, the main part of the external communication of companies relies on their websites. Assuming that corporate websites are written with the intention of highlighting the ‘best’ qualities of the firm, our intuition is that there will be a correalation between the web-based environmental culture indicator and the real environmental culture indicator developed by B-Lab.

2. Data and Methodology

2.1. Data

To create and test the tool we use two types of data:

1. The full-text of the companies’ websites that are B-Corp certified.
2. The B-Corp data.

B-Lab publicly releases the dataset with all the companies certified and the scores received. The scores include one main indicator, "overall score", which is an aggregation of five other indicators evaluating specific dimensions: governance, customers, workers, community, and environment. These dimensions are in turn divided into several items. In this paper, we focus solely on the B-Corp indicator concerning the "impact area environment" and as this is a pilot study we use only a subset of the B-Corp data limited to the Canadian and American companies. To create a corpus for each company, we identified URLs from the B-Corp data and downloaded text only from their homepages using the Wayback Machine. We used the Wayback Machine to download the pages as it was crucial to retrieve websites close to the certification date. We found 1741 company websites using the Wayback Machine. Next, we filtered the websites, choosing only English webpages and the most recent audit. Since a company can be certified more than once, we removed duplicates. Thus, the final sample has a total of 1110 firms, with 82% of companies from the US and 18% from Canada.

2.2. Methodology

Once the data has been prepared, the first step of the analysis consists in understanding the text of the corporate websites. Instead of counting specific keywords about predetermined topics, like most of the literature in social science, we use the Zero-shot text classification method which is a Natural language processing (NLP) task that is designed to answer the question: "Is this text about label X?" The answer to this question is an indicator of the confidence that the given text is about label X. The labels that we used for the purpose are the names of the items that compose the B-Corp environmental certification. Using the NLP model BART with the ZSTC, we aim to extrapolate the importance of a label. Then, our second task is simply calculating the Pearson correlations to measure which different items, among the ones included by B-Lab to evaluate the environmental impact certification, detected by BART in the text of the corporate websites are good proxies for the environmental score obtained by these firms.

The core of the tool is the Natural Language Processing (NLP) model "Bidirectional and Auto-Regressive Transformers" (BART)(Lewis et al. 2019), a transformed-based deep learning model for NLP developed by Facebook AI combining the most important characteristics of BERT and GPT. BART was pre-trained on English Wikipedia and BooksCorpus, using a two steps process: first, the text is changed by adding a noise factor (e.g., changing the words randomly), then, the model learns to reconstruct the original text. This new approach allowed BART to reach state-of-the-art performances in several NLP challenges.

We build the tool using BART on the Zero-shot text classification (ZSTC) task. ZSTC is a challenging task on the realms of the Natural Language Understanding problems, which

require the use of syntactic and semantic analysis to comprehend the actual meaning and sentiment of human language. More specifically, ZSTC refers to a task where the model classifies text into classes that were not present in the training corpus.

Performing the ZSTC requires choosing the labels and the corpus. Since we want to create a web-based environmental culture indicator, we use as labels the items that compose the 'impact area environment' index in B-Corp data (Table 1). After trying several settings, we decided to split each website into groups of 3 sentences to create our corpus. Indeed, we notice that the ZSTC performs better when the input is a text longer than a single sentence and smaller than the full website. Therefore, for each website, we perform the ZSTC on each group of sentences. It is important to highlight that the ZSTC produces a score among the several classes using cosine similarity metrics computation between the word-embedding vectors created by BART representing the label and the word embedding representation of the target corpus. The score is in a range from 0 to 1 and can be the same for more than one label. Then, to prepare the results for the Pearson correlation test, we take the average scores of each label for each website. In this way, for each website, we have the averaged results of the ZSTC for all the labels in Table 1 and the B-Corp data.

Table 1: Labels used in the Zero-shot text classification

• Air climate • Certification • Community • Construction practices • Designed to conserve agriculture process • Designed to conserve manufacturing process • Designed to conserve wholesale process • Energy water efficiency • Environment products services introduction • land office plant	• Environmental education information • Environmental management • Environmentally innovative agricultural process • Environmentally innovative manufacturing process • Environmentally innovative wholesal process • green investing • green lending • inputs • land life	• landwildlife conservation • material energy use • materials codes • outputs • renewable energy • cleaner burning energy • resource conservation • safety • toxin reduction remediation • training collaboration • transportation distribution suppliers • water

Source: https://data.world/blab/b-corp-impact-data/workspace/data-dictionary(2023)

3. Results

3.1. Zero-shot text classification

Table 2 shows a sample of the ZSTC results, which range from 0 to 1, representing the average score of each label for all the 1110 websites. Considering the mean score, "designed to conserve wholesale process", "inputs" and "safety" are the labels with the

higher average. In other words, in the full text of the website, on average there are more groups of sentences that, according to the model, refer to these labels. All labels have a minimum score of around 0, but the maximum values are not uniform. 17 out of the 31 labels have a maximum score over 0.90, meaning that at least once, each label is predominant on a website according to the model. On the other hand, the labels "clear burning energy," "green lending," "community," "designed to conserve manufacturing process," and "land wildlife conservation" have the lowest maximum value in the table, with the last label having a maximum score of less than 0.5. This suggests that our sample does not generally include websites where the "land wildlife conservation" label is more prominent than the other labels. Additionally, "land wildlife conservation," "environmentally innovative manufacturing process," and "green investing" have low average scores, indicating that the model rarely finds groups of sentences that correspond to these labels.

Generally, the minimum score value is closer to the mean than the maximum value, except for the scores of the labels "designed to conserve wholesale process" and "inputs." This suggests that while the other labels are frequently found on websites with low scores or not found at all, the model only considers the "designed to conserve wholesale process" and "inputs" labels when their scores are significantly higher than the others. Additionally, the scores in the table do not follow a normal distribution, as evidenced by the mean and median being unequal and the third quartile being closer to the minimum value, indicating possible outliers.

Table 2: Sample of ZSTC results averaged for all the companies

Labels	mean	std	min	25%	50%	75%	max
inputs	0.435	0.107	0.020	0.375	0.445	0.505	0.793
outputs	0.103	0.104	0.000	0.046	0.076	0.127	0.969
green investing	0.073	0.065	0.000	0.033	0.060	0.095	0.945
water	0.223	0.115	0.001	0.143	0.211	0.289	0.861
training collaboration	0.153	0.106	0.001	0.079	0.135	0.207	0.754

3.2. Correlation results

Table 3 shows the Pearson correlation results between each web-based environmental indicator and environmental indicator of B-Corp. As aforementioned, we ensure the normality of all the variables transforming them and testing skewness and kurtosis.

Table 3: Pearson Correlation results

Labels	r	p_value
Green investing**	0.497	0.000
Resource conservation**	0.472	0.000
Environmentally innovative wholesale process**	0.467	0.000
Green lending**	0.430	0.000
Environmental management**	0.390	0.000
Designed to conserve wholesale process**	0.356	0.000
Designed to conserve agriculture process**	0.349	0.000
Environmental education information**	0.320	0.000
Environmentally innovative manufacturing process***	0.297	0.000
Materials codes**	0.284	0.000
Designed to conserve manufacturing process**	0.283	0.000
Environment products services introduction**	0.266	0.000
Certification**	0.248	0.000
Environmentally innovative agricultural process***	0.215	0.000
Material energy use**	0.214	0.000
Outputs**	0.161	0.000
Land life**	0.108	0.000
Community*	0.081	0.007
Cleaner burning energy***	0.052	0.086
Renewable energy***	0.039	0.199
Inputs*	0.007	0.812
Air climate**	-0.003	0.922
Water***	-0.022	0.460
Safety**	-0.024	0.433
Land office plant**	-0.036	0.233
Toxin reduction remediation**	-0.067	0.025
Construction practices**	-0.079	0.008
Transportation distribution suppliers***	-0.080	0.008
Energy water efficiency**	-0.098	0.001
Training collaboration**	-0.161	0.000
Land wildlife conservation**	-0.220	0.000

Notes: The labels with * are transformed with the formula ln((label)+1)
The labesl with ** are transformed with the formula ln((label *10)+1)
The labels with *** are transformed with the formula ln((label *100)+1)

To ease the interpretation of the results we divide Table 3 in 3 parts. The lower box of the table contains the variables that have either negative or a null correlation with the B-Corp variable. Only the last 3 have p-values < 0,005 with the last two presenting p-value<0,001. Additionally, the last two labels have a score that is weakly inversely related. The label that are not in the two box are the one less important. The variables with a positive correlation

have a p-value less than 0.001, and three of them exhibit a correlation close to 0.30. However, when the correlation decreases and approaches zero, the p-value becomes insignificant. Finally, the most interesting is box on the highest part of Table 2. In this box, we find the labels that have the highest correlation with the environmental variable of B-Corp. They all have a p-value< 0,001 and a correlation higher than 0,30. Among these labels, the first four have a correlation higher than 0,4 with green investing that reaches almost 50% (0.497).

4. Conclusion

The goal of the research is to verify whether the ZSTC task can be used to create environmental indicators that are correlated with the real environmental indicators developed by B-Lab. Once we perform the ZSTC, we find significant correlations between most of the ZSTC scores and the environmental index measured by B-Lab. Specifically, we find that the scores of the topics: green investing, environmentally innovative wholesale processes, resource conservation, and green lending are the most correlated with the "impact area environment" indicator developed by B-Lab. This means that companies with a higher score on the environmental indicator created by B-Lab are likely to talk about the aforementioned topics.

Despite the promising results, our research presents three main limits. The first limit is inherent to the ZSTC task, which is considered one of the most challenging tasks for NLP models (Brown et al. 2020). This approach resembles an unsupervised method, which makes it generalizable. However, the model only has access to the label and the text, without any examples or further explanations, which forces it to interpret everything by itself. This can increase the misinterpretation and ambiguity of the already complicated natural language. Although this limit is intrinsic to the methodology, using a different state-of-the-art model instead of BART could potentially yield better results. It may be worth exploring other pre-trained language models such as LLAMA(Touvron et al. 2023) and PALM(Chowdhery et al. 2022) which have shown excellent performance in various NLP tasks. These models may have better capabilities to interpret complex natural language and reduce misinterpretation and ambiguity. The second limit is related to the labels used. We chose the labels directly from the items that B-Lab uses to evaluate the environmental culture of a company leaving to the model the interpretation of certain concepts. To overcome this limit, we could contact B Lab to obtain more appropriate labels that are specifically designed for the purpose of targeting certain environmental topics to reduce ambiguity and improve the accuracy of the results. Also, collaborating with domain experts, such as environmental scientists or sustainability practitioners, could also provide valuable insights for selecting appropriate labels. Finally, the third limit is connected to the correlation results. Indeed, the Pearson correlation partially explains the correlation

between the items and the environmental indicator. For the third limit, including control variables in a regression analysis can provide further insight into the relationship between the ZSTC score and the B-Lab variable, thereby enhancing our understanding of the observed correlation. Additionally, the regression analysis will allow us to predict the score that B-Lab can give to the companies based on their website text. For instance, it could be possible to examine companies that sought certification but were unable to obtain it, thus distinguishing between greenwashing and genuine green compliance.

References

Axenbeck, Janna, and Patrick Breithaupt. 2021. "Innovation Indicators Based on Firm Websites—Which Website Characteristics Predict Firm-Level Innovation Activity?" *PloS One* 16(4):e0249583.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33:1877–1901.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, and Sebastian Gehrmann. 2022. "Palm: Scaling Language Modeling with Pathways." *ArXiv Preprint ArXiv:2204.02311*.

Gök, Abdullah, Alec Waterworth, and Philip Shapira. 2015. "Use of Web Mining in Studying Innovation." *Scientometrics* 102(1):653–71.

Kinne, Jan, and David Lenz. 2021. "Predicting Innovative Firms Using Web Mining and Deep Learning." *PloS One* 16(4):e0249071.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. "Bart: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension." *ArXiv Preprint ArXiv:1910.13461*.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and Faisal Azhar. 2023. "Llama: Open and Efficient Foundation Language Models." *ArXiv Preprint ArXiv:2302.13971*.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16428

On the involvement of bots in promote-hit-and-run scams – the case of Rug Pulls

Dietmar Janetzko[1], Jonas Krauß[1], Frederic Haase[2], Oliver Rath[2]
[1]Stockpulse GmbH, Germany, [2]University of Cologne, Germany

Abstract

Many social media frauds related to finance can be summarized under what we consider promote-hit-and-run scams. Examples include rug pull scams also known as exit scams, pump-and-dump schemes or bogus crypto currency trading platforms. For scams of this kind *to work they must be publicly advertised as lucrative investment opportunities. Social media are key in this promotion. Here, fraudsters find platforms to persuade others investing into what later turns out to be a scam. Via social network analysis of Twitter screen names and their first-level contacts, our work investigates rug pulls. It examines social media communication around them with a special focus on the deployment of bots. Repeatedly bots have been identified in social media campaigns (Orabi et al., 2020). Bot deployment in the context of rug pulls, however, has not been studied yet. Our analysis of social data of 27 rug pulls reveals for the first time massive bot activity coordinated within and between rug pulls mainly targeting established finance news outlets, e.g., Bloomberg, Reuters. Among the conclusions of our work is that bot deployment may prove an early indicator for rug pulls and other promote-hit-and-run scams.*

Keywords: *Bots; social media; rug pull; crypto currencies; fraud; twitter*

The work described in this paper has been partially financed by the AFFIN Project funded by the German Federal Ministry of Education and Research (BMBF).

1. Introduction

Financial fraud in social media comes in various forms (e.g., Mirtaherie et al, 2021 Nghiemet et al. 2021). Many of them share a similar pattern: fraudsters entice people into actions that promise high gains, but which eventually translate into losses to the victims. We summarize swindles implemented around this pattern under the category of promote-hit-and-run scams. Examples include rug pull scams, pump-and-dump schemes or enticements to bogus crypto currency trading platforms. Most of them are initiated by coordinated social media campaigns on platforms such a Telegram, Reddit or Twitter (Orabi et al., 2020; Sharma et al, 2021; Tardelli et al., 2020). The digital traces fraudulent campaigning leaves gives us a handle to identify them. The work in this paper investigates Twitter activities related to rug pulls (Solidus, 2022, Scharfman, 2023). Rug pulls are examples of promote-hit-and-run scams usually connected to crypto currencies whereby the developers take the assets of investor and disappear. Negligible before 2020, rug pulls now spread at a staggering speed. Platforms like Twitter, Reddit or Telegram are among the most popular communication channels for rug pull promotion. The importance of communication for the success of rug pulls invites a comparison with other campaigns that rely on communication on social media, e.g., pump-and-dump (Mirtaheri et al., 2021) or in politics (Murthy et al., 2016). In those and many other areas communication on social media is often affected or even orchestrated by bots. To the best of our knowledge, however, possible bot involvement of social media communication related to rug pulls has not yet been investigated.

In sum, public communication is indispensable to promote rug pull projects. Details on the involvement of bots as an essential element of this type of fraudulent communication are largely unknown, however. Against the background of this research gap, the overall question addressed in this paper is: Are there social media indicators of bots activity related to crypto investments that were later rug pulled? We break this question down into two more specific objectives: Firstly, to find out whether and in which way bot activity is reflected in historical social data of rug pulls, we compare screen names co-occurrence across 27 of them. We secondly identify screen names involving a rug pull and their contacts and set up a social network with screen names as vertices and interactions (replies, retweets) between them as directed edges. We then examine whether social network metrics like in- and out-degree centrality and their relation to bot probability scores (Yang et al, 2020) allow us to identify the deployment of bots in rug pulls. Our paper is organized as follows. The next section looks into recent work on bot activity around finance-related topics in social media, which is followed by an outline of the methodology used. Next, the studies corresponding to our research objectives are presented. The paper concludes with a discussion of our results.

2. Related Work

Social Bots (or simply bots) are automated programs that deceitfully act like humans on social media. Though bots are a widespread phenomenon (Orabi et al., 2020, Aljabri et al., 2023) there are only a few studies on bots in relation to crypto currencies and to their role in financial markets in general (Mirtaheri et al., 2021; Schuchard et al 2019; Tardelli et al., 2020). Of particular interest to our work is the work by Mirtaheri et al., 2021. The authors found a large number of Tweets related to crypto currencies generated by bots. Increased Twitter bot activity was observed during the time of pump-and-dump frauds. Though pump-and-dump schemes share some similarities with the rug pulls, they are not necessarily the same. For instance, a rug pull involves completely new tokens. Thus, it can be expected that in rug pulls more effort has to be put into the trust building. This can hardly work in a short time span and may require a strong visibility on social media. It cannot be ruled out that various types of bots are deployed to achieve these goals. Examples include the follower-bots that increase the sheer number of accounts following another account or commenting bots that post comments.

3. Method

Data on the rug pulls studied was made available by Stockpulse's archive of historical and current social media messages.[1] It covers historical and current social media communication analytics for, e.g., Twitter, Telegram, Reddit, Discord over up to 12 years. From Stockpulse's archive, Tweet data on each of the following 27 rug pulls was analyzed:

Africrypt, AnubisDAO, Baby Musk Coin, BabyEth, BankSocial, Billionaire Dogs, BitConnect, CryptoZoo, DeFi100, Dink Doink, EthereumMax,Freeway Token,Green Satoshi, Kronos DAO, MILF Token, Mango Token, RapDoge, SafeMoon V1, Snowdog DAO, Snowflake DeFi, SolFire, Squid Game, StableMagnet, TeddyDoge, Terra Classic, Thodex, Yummy Token

For our studies we selected medium to large rug pulls executed between Feb 15, 2016 – Feb 6, 2023. The time of the communication sourced was essential to increase the likelihood of catching bot activity. Therefore, we collected Tweets exchanged after the introduction of the coin to the public associated with each rug pull examined, but before its public exposure

[1] Stockpulse is a data analytics company based in Bonn (Germany) offering advanced AI-driven data and signals for financial institutions and regulators.

as a scam via an authoritative source that flagged up the fraud.[2] The data-collection is broken down into two steps:

Step 1. Using the interval-based method described above we sourced messages referring to one of the 27 rug pulls and preprocessed them into key variables including screen name and time. This resulted in 27 seed lists of screen names, which is just a list of Twitter addresses of account connected to a rug pull. The 27 seed lists ranged between 23 (Snowdog DAO) and 200,566 (Squid Game) unique screen names.

Step 2. For each screen name of the 27 seed lists, a list of its friends (contacts) was sourced with 2 screen names (source, target) per data row. Again, 27 lists resulted. In contrast to the seed lists of step 1, the contact lists qualify as relational data as it reflects communication links between source-target pairs of screen names.

4. Studies and Results

Study 1 proceeds on the assumption that Twitter screen names overlap across otherwise unrelated rug pulls suggesting that the same group of bots are being deployed across different rug pulls. To examine screen name overlap we considered all 351[3] pairwise intersections of screen names that could be generated from the 27 rug pulls studied.

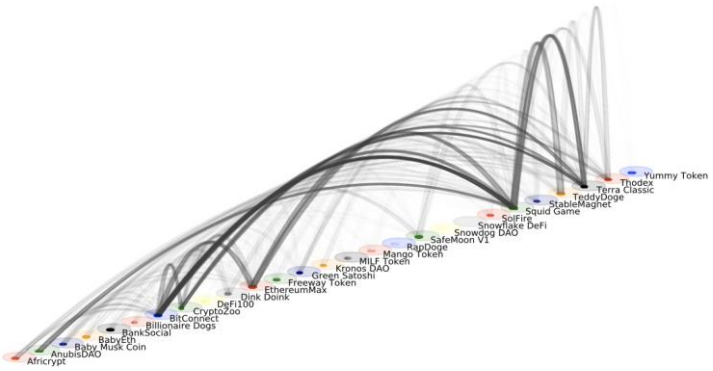


Figure 1. Screen Name Co-Occurrence across Rug Pulls

Fig. 1 visualizes screen name co-occurrence between different rug pulls. It highlights that in a number of rug pulls the same screen names occur. Differences in thickness of the bended

[2] Authoritative sources include government agencies, private or public regulating bodies, established news outlets or companies focused on crypto security or trade surveillance.

[3] (27*27-27) / 2 = 351

connecting lines expresses that rug pulls are affected to a different extent by screen name co-occurrence. Squid Game, Bitconnect, and Terra Classic stand out here. They are rug pulls that show high co-occurrence values in pairwise rug pull comparisons (thickness of bended connecting lines). They also come with the highest number of co-occurrences with other rug pulls (number of bended connecting lines).

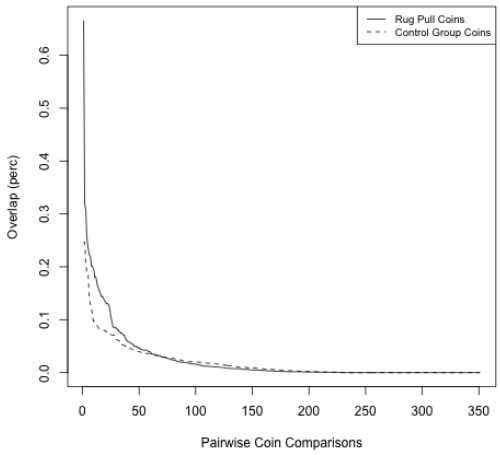


Figure 2. Screen Name Co-Occurrence across Rug Pull and Coins of a Control Group

Fig. 2 shows the screen name overlap across 27 rug pulls and across 27 coins of a control group. The figure plots on the x-axis all pairs of coins against the relative frequency of the screen name co-occurrence found for each of them. The resulting negative exponential distribution (solid line) reveals that rug pull coins share up to 66% of screen names (and possibly participants) with each other. At the same time, the long tail of the distribution suggests that many rug-pulls seem to be largely disjoint in terms of screen name overlap. The dashed line shows the comparison values of a control group of crypto currencies[4]. Here, the screen name co-occurrence ranges between 0.09% and 26%. Testing the differences between both distributions with a two-sample Kolmogorov-Smirnov test revealed a significant difference ($D=.160$, $p<0.001$). Clearly, high proportions of identical screen names showing up in pairwise comparisons of rug pulls is suspicious. But in itself it is not hard evidence for bot activity. The results of study 1 alone cannot rule out alternative

[4] The control group of coins was made up of the top 27 coins found on coinmarketcap.com on Feb 10, 2023 (ADA,APT,ATOM,AVAX,BCH,BUSD,CRO,DAI,DOGE,DOT,ETC,FIL,HBAR,LDO,LEO,LINK,LTC,MATIC,OKB,SHIB,SOL,TON,TRX,TUSD,UNI,WBTC,XLM,XMR,XRP). The top three coins (BTC, ETH, USDT) were not included as the sheer volume of those coins made them unlikely matches for the rug pull coins studied.

explanations of the findings, e.g., that humans and not bots have created the screen name overlap spotted.

Study 2: The open questions of study 1 prompted a second study to find out more about overlapping screen names leveraging social network analysis (SNA). We examined the 25 screen names (Twitter accounts) with the highest in-degree and out-degree centrality, respectively, found in three largest rug pulls examined.[5] In our study, the out-centrality for a screen name reflects the number of retweets or replies actively *sent*. By contrast, the in-degree centrality is a measure of the number of retweets or replies that a screen *received*.

Table 1: Mean Bot Probability of Top 25 Senders and Receivers for 3 Major Rug Pulls

Rug Pull	**messages**	**mean *p* in-centrality**	**mean *p* out-centrality**
Squid Game	200,566	0.102	0.863
Terry Classic	36,062	0.296	0.581
Bitconnect	11,463	0.122	0.901

Table 1 has results of study 2 that connected centrality scores with bot probabilities. For each of the 3 × 2 × 25 screen names we collected its Botometer score, viz., its probability of being a bot (Yang et al, 2020). For the 25 nodes with the highest in- and out-degree centrality we averaged this score. We found the mean bot probability (Botometer Score) to be consistently higher for screen names with high out-degree centrality and vice versa. The findings offer answers to the questions that remained open after study 1. They suggest that in the networks analyzed the communication is mainly driven by bots and not by humans (Fig. 3. left) and that the bots are not non-tweeting follower-bots as they are actively targeting popular accounts, e.g., those of influential news outlets (Fig. 3. right).

[5]In-degree and out degree centrality are metrics provided by social network analysis. A node in directed graph on which many other nodes are pointing has a high in-degree centrality. In social network based on Twitter data such a node is a typical receiver. Vice versa, a node in directed graph that points massively onto other nodes has a high out-degree centrality. In an actor network based on Twitter data such a node is a typical sender.

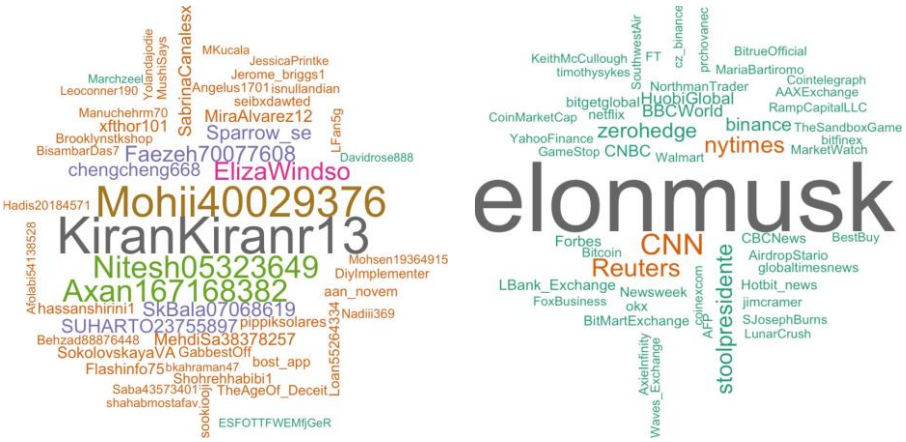


Figure 3. Twitter Screen Names with high out-degree Centrality (left) and with high in-degree Centrality (right)

5. Discussion

Bots are strongly involved in rug pulls and our explorative study identified distinctive patterns of bot deployment both within a single and also across several rug pulls. Furthermore, out work suggests that bot activity is significantly stronger in rug pulls than in other crypto currencies. *Within one rug pull*, we found strong evidence that bots drive the communication by the sheer number of messages and their specific targeting. The large number of messages sent out was reflected by a high out-degree centrality. Their specificity became evident when we looked at the receiving Twitter accounts. Here, we often find popular sites, many of them news outlets, e.g., Reuters, CNN or Bloomberg. It is tempting to conjecture that bots may repeatedly reply to Tweets of an established site in order to build up trust simply by what is known in psychology as a mere-exposure effect (Zajon, 1968). *Bot deployment across rug pulls* could be identified via screen name co-occurrence identifiable across multiple rug pulls. Together with aforementioned results it suggests that often the same bots are being used across rug pulls. This finding raises questions and concerns. It obviously reflects professional coordinated efforts to defraud investors possibly via bot net services. We did not study the actual effects bots have on humans. But it can be reasonably expected that bot activities identified in our studies create a false sense of interest in a project, making it difficult for investors to distinguish between genuine interest in a project and artificially generated buzz/hype. In our analysis we harnessed historical data. Follow up research is required to find out whether bot activity patterns can actually predict future rug pulls. Further research is required to study those effects and whether the patterns found also characterize other promote-hit-and-run scams.

References

Aljabri, M., Zagrouba, R., Shaahid, A., Alnasser, F., Saleh, A., & Alomari, D. M. (2023). Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining*, *13*(1), 20.

Mirtaheri, M., Abu-El-Haija, S., Morstatter, F., Ver Steeg, G., & Galstyan, A. (2021). Identifying and analyzing cryptocurrency manipulations in social media. *IEEE Transactions on Computational Social Systems*, 8(3), 607–617.

Murthy, D., Powell, A. B., Tinati, R., Anstead, N., Carr, L., Halford, S. J., & Weal, M. (2016). Automation, algorithms, and politics| Bots and political influence: A sociotechnical investigation of social network capital. *International Journal of Communication*, *10*, 20.

Nghiem, H., Muric, G., Morstatter, F., & Ferrara, E. (2021). Detecting cryptocurrency pump-and-dump frauds using market and social signals. *Expert Systems with Applications*, 182, 115284.

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: a systematic review. Information Processing & Management, 57(4), 102250

Scharfman, J. (2023). *The Cryptocurrency and Digital Asset Fraud Casebook*. Springer Nature.

Schuchard, R., Crooks, A., Stefanidis, A., & Croitoru, A. (2019). Bots in nets: empirical comparative analysis of bot evidence in social networks. In *Complex Networks and Their Applications VII: Volume 2 Proceedings The 7th International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2018 7* (pp. 424-436). Springer International Publishing.

Solidus, I. (2022). The Rug Pull Report. New York. Retrieved from https://www.soliduslabs.com/reports/rug-pull-report

Tardelli, S., Avvenuti, M., Tesconi, M., & Cresci, S. (2020). Characterizing social bots spreading financial disinformation. In G. Meiselwitz (Ed.), Social computing and social media. design, ethics, user behavior, and social network analysis (pp. 376–392).

Yang, K. C., Varol, O., Hui, P. M., & Menczer, F. (2020, April). Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 01, pp. 1096-1103).

Zajonc, R. B. (1968). Attitudinal Effects of Mere Exposure. Journal of Personality and Social Psychology, 9, 2, 1–27.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16470

Exploring expert opinion on climate policy using Twitter

Enrico Bergamini[1], Ivan Savin[2], Jeroen van den Bergh[2,3,4]
[1]Department of Economics and Statistics "Cognetti de Martiis", Università di Torino, Italy, [2] Institute of Environmental Science and Technology, Universitat Autònoma de Barcelona, Spain, [3] ICREA, Barcelona, Spain, [4] School of Business and Economics & Institute for Environmental Studies, Vrije Universiteit Amsterdam, The Netherlands.

Abstract

We study online conversations about climate policy by building a novel dataset of around 100,000 tweets and tweet threads by climate policy scientists. This data is complemented with information about the scientific affiliation and production of scientists. We undertake an exploratory analysis of the content of tweets by means of Natural Language Processing. In addition, we study the relationship between tweet content and academic background. This indicates that economists and political scientists are the most active in discussing climate policy on Twitter. We further find that the policy instruments receiving most attention are cap-and-trade and carbon taxation.

Keywords: *climate policy; carbon price; scientometrics; social media; topic modelling.*

1. Introduction

Insights and views of scientists nowadays appear not only in academic publications and popular science but increasingly in social media. This also holds true for the theme of climate policy. In this study we examine conversations on Twitter, the most important platform for scientists, and match this to data on academic production. Scientists' presence and influence on social media has received considerable attention in recent years. Researchers have many incentives to use social media: getting informed or informing others about recent data, findings and publications, staying updated about planned conferences and workshops, debating topical issues in political reality or themes in their scientific field, and exchanging ideas with journalists, policymakers, environmental NGOs and the general public (Howoldt et al. 2023).

Many scientists choose Twitter over other social media, which affirms its role as a medium for experts, journalists and politicians to meet and engage in public conversation (Della Giusta et al.; Greetham 2021). For these reasons, Twitter has proven to be a good tool for exploring public and élite opinion about relevant phenomena (Bollen et al. 2011). Recent efforts have quantified behaviour and characteristics of scholars with social media data, raising opportunities for new instruments in scientometrics (Sugimoto et al. 2017a). Twitter has been used as a resource to study quantitatively the interaction between scientific production and social media uptake and engagement of academics (Howoldt et al. 2023). Next to promoting scientific work within the community and popularise it among a broader audience, scientists engage and learn from their communities and peers. An emerging literature is studying scientific sub-communities on Twitter. Côté and Darling (2018) found that most Twitter interactions by scientists are directed at other scientists. Della Giusta et al. (2021) compare the behaviour and communication of top economists vis-à-vis natural scientists. They find that the communities tend to behave in different ways: while economists explain more and engage less, natural scientists care more about communicating with the general public. Another study by Khandelwal and Tagat (2021) instead examines the communication of development researchers by combining Twitter and survey data. Bisbee et al. (2022) study the network of political scientists on Twitter with a focus on the United States. All these studies show that observing online networks and conversations can help provide evidence on knowledge diffusion as well as on different characteristics of scientific communication (Alperin et al. 2019; Howoldt et al. 2023).

However, an open challenge to studying scientific interactions on social media is a precise identification of accounts belonging to scholars. Recent work (Mongeon et al. 2022) addresses this gap, making a large-scale dataset of academics available. This allowed connecting academics from a large range of fields to their social media accounts. Researchers have employed Twitter data to track public opinion on a variety of climate issues: from polarization (Jang and Hart 2015) through misinformation (for a review see Treen et al. 2020) to social movements (Chen et al. 2022; Thorson and Wang 2020) and COP meetings (e.g.

Pearce et al. 2014; Hopke and Hestres 2018; Sanford et al. 2021). Jang and Hart (2015) study the polarization in climate change narratives among the general public by employing big data from Twitter. Cody et al. (2015) analysed public opinion about climate change, linked to events like climate disasters and legislation. Veltri and Atanasova (2017) studied the climate-change discourse on Twitter, mapping a sophisticated and complex information ecosystem around climate change, more nuanced than other studies would suggest. Another study investigates climate policy debates: Wei et al. (2021) explore the networks of accounts and conversations about the European Union's Emissions Trading System (EU-ETS), finding a prominent role of government officials and industry practitioners, and a focus on policies, legislations, prices and allocation.

While many scholars have put great effort in collecting (big) datasets (Effrosynidis et al. 2022) and studying climate change public opinion on Twitter, other studies have focused on specific subsets of users. Vu et al. (2020) studied the networks of climate NGOs on Twitter, highlighting the role of climate opinion leaders. They quantify the importance of network centrality in opinion leadership and suggest a strong Global North versus South division. Almironet al. (2022) study the network of think-tanks with contrarian stances on climate change in Europe, and their ties to the United States. Goritz et al. (2022) study the online presence in terms of climate policy of International Organizations, also employing Twitter data. Walteret al. (2019), instead, focus on scientists and climate change communication using network analysis on Tweets. They provide fascinating evidence on the varying communicative strategies of scientists when debating with different types of accounts (journalists, politicians, other scientists).

In this study, we focus the attention on opinions of scientists about climate policies. Earlier, Drews et al. (2023) and Savin et al. (2023) conducted an online survey among researchers who published on the topic of climate policies in the last five years finding that direct regulation is the most favoured type of instrument in a policy mix, while carbon tax and carbon market face more resistance from scientific field like political science, agriculture and sustainability transition. However, this study had a low response rate (less than 5%) and could not assure that all disciplines were properly covered. In the present paper we aim to expand this earlier work by analysing a much larger sample of scientific experts on climate policy, to verify previous conclusions and obtain additional insights about the main differences among scientists regarding opinions about climate policy.

We study the content of tweets using Natural Language Processing (NLP). To our knowledge, no study has comprehensively mapped scientific communication on social media specifically looking at climate policies. We fill this gap by collecting a large dataset of tweets about climate policies written by scientists. We add to two streams of literature: the one studying scientists' social media presence, and the other exploring climate discourse on social media. Our motivation stems from the importance of understanding scientific communication

and consensus as well as potential controversies about climate policies. Understanding expert opinions on climate policy may provide additional evidence for the support of decarbonisation policies (Drews and van den Bergh 2016). Furthermore, we add evidence and nuance by linking information about academic characteristics to tweets about climate policies.

2. Data and method

To answer the questions posed in the previous section, we build a novel database of Tweets around climate policies. We isolate specifically the subset of tweets made by scientists. We start by collecting from Tweets from the Twitter Academic API, corresponding to the keyword search based on Drews et al. (2023). This set of keywords identifies climate policy by focusing on a set of known instruments (carbon taxing, cap-and-trade, etc). The resulting database comprises 9.2 million tweets from 1.5 million accounts during the period 2007 to 2022. To assess which accounts belong to scientists, we rely on Mongeon et al. (2022) who created an algorithm to match known databases of scientists from all disciplines with Twitter accounts. This resulted in an open-access database of around 500,000 scientists matched with Twitter account ID's. In order to complement this information, we reconstruct the full database used in their paper by querying OpenAlex for information about scientists' main scientific field, affiliation, and academic performance metrics (e.g. number of publications and citations).

We merge the two databases, obtaining information about which tweets belonged to scientists in the Twitter database. This indicates they contributed to 4% of total tweets, namely 360,000 tweets out of 9.2 million tweets in total (i.e. corresponding to the search query). Note that this excludes non-English tweets, retweets and replies to other users from the sample. Threads are also a popular instrument on Twitter. They are chained tweets to overcome the character limit of a single tweet (280 characters). They allow for a more elaborate explanation of a scientific idea. Therefore, we include them and treat them as if they were single extended tweets. In terms of users, we find around 13,000 unique Twitter accounts belonging to scientists that we can match to the database of Mongeon et al. (2022).

3. Preliminary results

3.1 Descriptive statistics

The resulting dataset, cleaned for non-English tweets and replies, comprises 71620 tweets and 8565 threads, resulting in a total of 80185 unique observations, written by 13093 scientists. Table 1 shows the volume and percentage of Tweets by main academic discipline. The breakdown of disciplines follows that of OpenAlex which is based on Wikidata ontology (see also Mongeon et al. 2022).

The table shows that economists tweet relatively much compared to other scientists, accounting for a fifth of the total number of tweets. Economists also write relatively many threats.

Table 1. Distribution of tweets, accounts and threads by discipline

Field	**Tweets**		**Scientists**		**Threads**	
	volume	**%**	**number**	**%**	**volume**	**%**
Political science	15792	19.69	2445	18.63	1331	15.54
Economics	15697	19.58	1323	10.08	2164	25.27
Biology	15277	19.05	2900	22.10	1367	15.96
Non classified	8232	10.27	1153	8.79	963	11.24
Psychology	3405	4.25	681	5.19	528	6.16
Physics	2912	3.63	467	3.56	469	5.48
Computer science	2828	3.53	738	5.62	238	2.78
Business	2709	3.38	468	3.57	222	2.59
Environmental science	2390	2.98	350	2.67	289	3.37
Medicine	2358	2.94	729	5.55	160	1.87
Geology	2250	2.81	329	2.51	177	2.07
Engineering	1926	2.40	247	1.88	347	4.05
Geography	1561	1.95	382	2.91	84	0.98
Philosophy	760	0.95	189	1.44	98	1.14
Sociology	563	0.70	182	1.39	15	0.18
Mathematics	529	0.66	142	1.08	54	0.63
Chemistry	344	0.43	153	1.17	15	0.18
History	338	0.42	128	0.98	29	0.34
Art	250	0.31	87	0.66	8	0.09
Materials science	64	0.08	31	0.24	7	0.08

To explore the content of tweets, we perform a cleaning and pre-processing as is common in studies using NLP. In order to further explore the dataset, we analyse the content of tweets. We rely on BERT language models based on word embeddings, as proposed by Grootendorst (2022). Topic modelling uses word distributions across documents to extract latent topics for each document. We estimate a baseline topic model to cluster scientists' tweets about climate policies. Figure 1 shows a semantic map of the topics derived. Each tweet is a dot on the map, while the coloured clusters are topics. Although the total number of topics is 93, the figure depicts labels for the top 25 topics in terms of overall volume. Visual inspection indicates a prevalence of two instruments in the debate, namely carbon taxation and cap-and-trade. Due to a focus on English-only tweets, there is a bias towards national debates in Australia, Canada, and the United States. A specific cluster about EU climate policies and the EU's emissions trading system (ETS) appears as well, which is in line with the results presented by Wei et al. (2021). Fossil fuel subsidies (10) is also a large topic, as is the general debate on energy (13). Smaller topics include trade-offs between climate policy and

economic growth (15), climate policies for housing and cities (14), and policies linked to forests (17). The evidence provided in this paper is still preliminary. Nevertheless, our results show that it is possible to identify debates on social media by scientists using large data. In a next phase we will connect the novel dataset to information about scientists' academic characteristics and activities to explore topic heterogeneity. Further analysis could exploit Structural Topic Models in order to estimate covariates for topic prevalence and formation (e.g., in terms of academic field). Content, sentiment, network and psychometric analyses will provide further insights into climate policies communication.

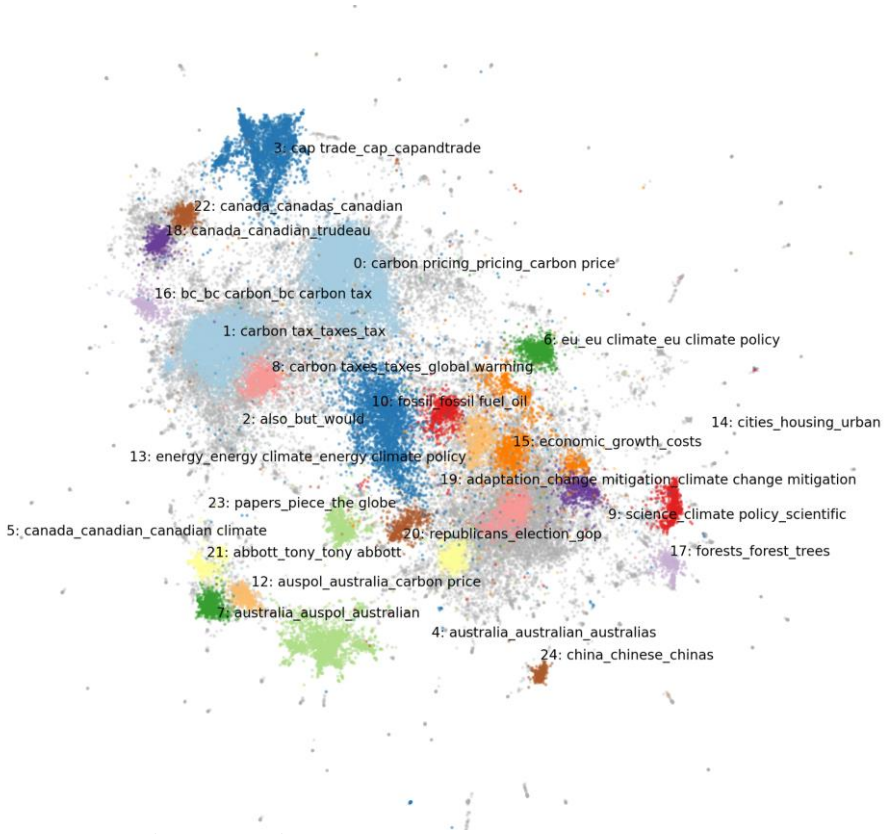


Figure 1. A semantic map of main topics of tweets by scientists on climate policy. Source: Authors elaboration

References

Almiron, N., Moreno, J. A., & Farrell, J. (2022). Climate change contrarian think tanks in Europe: A network analysis. *Public Understanding of Science*, 09636625221137815.

Alperin, J. P., Gomez, C. J., & Haustein, S. (2019). Identifying diffusion patterns of research articles on Twitter: A case study of online engagement with open access articles. *Public Understanding of Science*, 28(1), 2-18.

Bisbee, J., Larson, J., & Munger, K. (2022). #Polisci Twitter: A descriptive analysis of how political scientists use Twitter in 2019. *Perspectives on Politics*, 20(3), 879-900.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *In Proceedings of the International AAAI Conference on Web and Social Media*, 5, 450-453.

Chen, K., Molder, A. L., Duan, Z., Boulianne, S., Eckart, C., Mallari, P., & Yang, D. (2022). How climate movement actors and news media frame climate change and strike: Evidence from analyzing Twitter and news media discourse from 2018 to 2021. *The International Journal of Press/Politics*, 19401612221106405.

Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2015). Climate change sentiment on Twitter: An unsolicited public opinion poll. *PloS One*, 10(8), e0136092.

Côté, I. M., & Darling, E. S. (2018). Scientists on Twitter: Preaching to the choir or singing from the rooftops? *Facets*, 3(1), 682-694.

Della Giusta, M., Jaworska, S., & Vukadinović Greetham, D. (2021). Expert communication on Twitter: Comparing economists' and scientists' social networks, topics, and communicative styles. *Public Understanding of Science*, 30(1), 75-90.

Drews, S., Savin, I., & van den Bergh, J. C. (2023). A Global Survey of Scientific Consensus and Controversy on Climate Policy (*mimeo*)

Drews, S., & van den Bergh, J. C. (2016). What explains public support for climate policies? A review of empirical and experimental studies. *Climate Policy*, 16(7), 855-876.

Effrosynidis, D., Karasakalidis, A. I., Sylaios, G., & Arampatzis, A. (2022). The climate change Twitter dataset. *Expert Systems with Applications*, 204, 117541.

Goritz, A., Schuster, J., Jörgens, H., & Kolleck, N. (2022). International public administrations on Twitter: A comparison of digital authority in global climate policy. *Journal of Comparative Policy Analysis: Research and Practice*, 24(3), 271-295.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint* arXiv:2203.05794.

Hopke, J. E., & Hestres, L. E. (2018). Visualizing the Paris Climate Talks on Twitter: Media and climate stakeholder visual social media during COP21. *Social Media*

Howoldt, D., Kroll, H., Neuhäusler, P., & Feidenheimer, A. (2023). Understanding researchers' Twitter uptake, activity and popularity—an analysis of applied research in Germany. *Scientometrics*, 128(1), 325-344.

Ivanova, A., Schäfer, M. S., Schlichting, I., & Schmidt, A. (2013). Is there a medialization of climate science? Results from a survey of German climate scientists. *Science Communication*, 35(5), 626-653.

Jang, S. M., & Hart, P. S. (2015). Polarized frames on 'climate change' and 'global warming' across countries and states: Evidence from Twitter big data. *Global Environmental Change*, 32, 11-17.

Khandelwal, A., & Tagat, A. (2021). #DevResearch: Exploring development researchers' Twitter use for research dissemination. *Scholarly and Research Communication*, 12(1), 23-pp.

Mongeon, P., Bowman, T. D., & Costas, R. (2022). An open dataset of scholars on Twitter. *arXiv preprint* arXiv:2208.11065.

Pearce, W., Holmberg, K., Hellsten, I., & Nerlich, B. (2014). Climate change on Twitter: Topics, communities and conversations about the 2013 IPCC Working Group 1 report. *PloS One*, 9(4), e94785

Sanford, M., Painter, J., Yasseri, T., & Lorimer, J. (2021). Controversy around climate change reports: A case study of Twitter responses to the 2019 IPCC report on land. *Climatic Change*, 167(3-4), 59.

Savin, I., Drews, S., & van den Bergh, J. C. (2023). Carbon Pricing: Perceived Strengths, Weaknesses and Knowledge Gaps according to a Global Expert Survey. (*mimeo*)

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, 68(9), 2037-2062.

Thorson, K., & Wang, L. (2020). Committed participation or flashes of action? Mobilizing public attention to climate on Twitter, 2011–2015. *Environmental Communication*, 14(3), 347-363.

Treen, K. M. d'I., Williams, H. T. P., & O'Neill, S. J. (2020). Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5), e665.

Veltri, G. A., & Atanasova, D. (2017). Climate change on Twitter: Content, media ecology and information sharing behaviour. *Public Understanding of Science*, 26(6), 721-737.

Vu, H. T., Do, H. V., Seo, H., & Liu, Y. (2020). Who leads the conversation on climate change?: A study of a global network of NGOs on Twitter. *Environmental Communication*, 14(4), 450-464.

Walter, S., Lörcher, I., & Brüggemann, M. (2019). Scientific Networks on Twitter: Analyzing Scientists' Interactions in the Climate Change Debate. *Public Understanding of Science*, 28(6), 696–712.

Wei, Y., Gong, P., Zhang, J., & Wang, L. (2021). Exploring Public Opinions on Climate Change Policy in "Big Data Era"—a Case Study of the European Union Emission Trading System (EU-ETS) Based on Twitter. *Energy Policy*, 158, 112559.

Networks and narratives on Twitter about the #8M International Women's Day (2018) in Spain: Feminist Social Movement and counter-movement expressions

Elena Ruiz-Angel[1], Patricia Ruiz-Angel[2], Francisco Javier Santos[1], Estrella Gualda[1]
[1]Social Studies and Social Intervention Research Center (ESEIS), Center for Research in Contemporary Thought and Innovation for Social Development (COIDESO), Department of Sociology, Social Work and Public Health, University of Huelva, Spain. [2]Department of Sociology, Social Work and Public Health, University of Huelva and Department of Sociology University Pablo de Olavide, Spain.

Abstract

On March 8, 2018, International Women's Day took place worldwide, which brought relevant mobilisations and support in Spain. The feminist movement proved strong and demonstrated great vitality in a historic and unprecedented mobilisation. That day, many people took to the streets worldwide, and massively in Spain, to demand equal rights and opportunities for women and men. This mobilisation also took place on social networks. This paper aims to analyse the networks and narratives on Twitter around March 8 virtual mobilisation in Spain in 2018. This work analyses 557,548 tweets containing the hashtags representative of the mobilisation and collected through the API rest and API streaming Twitter platforms. The results suggest the presence of a strong national and international network of support for the feminist movement and a counter-feminist network that does not support the mobilisation and also propagates hate speech towards women and the feminist movement itself on the Twitter network.

Keywords: *Social Networks; Semantic Networks; Narratives; Hate Speech Online; International Women's Day; Twitter.*

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16439

Suitable statistical approaches for novel policies: spatial clusters of childcare's services in Veneto, Italy

Angela Andreella[1], Stefano Campostrini[1]
[1]Department of Economics, Ca' Foscari University of Venice, Italy

Abstract

More and more often, policymakers face complex problems that require suitable information obtainable only from the "intelligence of data." This can be obtained by analyzing several data sets (many of high dimension) and adopting suitable, often "sophisticated," statistical models. Here we deal with policies for affordable and quality childcare, essential to balance work and family life, increase labor market participation, promote gender equality, and fight against fertility decline. Understanding the complex dynamics of demand and supply of childcare services is challenging due to the nature of the data: high-dimensional, complex, and heterogeneous nationwide. Considering the Italian case, this complexity and heterogeneity are partially due to the lack of governance at the regional level leading to immediate and effective new policies challenging. This paper aims to analyze the multidimensional aspect of the supply-demand of childcare services combination in the Veneto Italian region using a novel statistical approach and an innovative dataset. We apply the regionalization approach (a clustering method with spatial constraints) to give an immediate picture of childcare services' supply and demand variability. Our empirical findings confirm how the Veneto region is described by many "sub-regional models," providing a preliminary attempt to demonstrate how socio-demographic factors drive these patterns.

Keywords: *clustering; childcare services; supply and demand; social services; spatial proximity.*

1. Introduction

In recent years, there has been an increased focus on analyzing the social impact of childcare services. In Italy, such services were established in 1971 as a "social service in the public interest," while in the late 1990s, additional early childhood services were introduced in Italy. These services were established as social benefits whose primary purpose was to support early childhood and parents, especially women, to care for their children and participate in the labor market. There is a vast literature regarding the effects of childcare services policies on female labor market participation (Landivar *et al*., 2021, Borghorst *et al.* 2021), fertility (Del Boca *et al*., 2002), and cognitive child improvements (Brilli *et al*., 2016). Therefore, understanding the status of childcare services in the territory as well as the socio-demographic characteristic of the population (Plantega *et al*., 2009) is essential to support policymakers in ideating and applying immediate and effective policies.

Unfortunately, Italy is characterized by a low child coverage rate (i.e., the number of places available in the childcare structure divided by the number of children under two ages) and by a low female labor force participation rate (Dipartimento per le politiche della famiglia, *et al*., (2020)). This situation has sparked a heated debate about the role of these services, and there is a consensus that the education system for children must be improved (Dipartimento per le politiche della famiglia, *et al*., (2020)). From a practical point of view, there are difficulties for the supply system to respond quickly to the changes in the sociocultural framework and the specific needs for quality educational services nationwide. First, structural deficits in services have emerged despite the potentially great demand. Second, the spatial distribution of childcare services is highly irregular within and between the Italian regions. Creating new policies on the ground in Italy is complex and challenging.

Motivated by these questions, this paper aims to provide an approach and methods to analyze the complex high-dimensional structure of the combination of supply and demand of childcare services. We want to highlight how the supply and demand of childcare services do not follow "regional patterns" but are affected by sub-regional variability driven by socio-demographic factors. In order to find these "sub-regional models," a suitable statistical approach must be considered that can extract information from different socio-economic variables taking into account the spatial geographical context.

The "sub-regional models" we want to find and analyze can be seen as clusters of administrative units (ATS – Ambiti Territoriali Sociali: aggregation of municipalities specifically created to plan and provide social services for the local population) sharing a similar combination of childcare services' supply and demand characteristics. The regionalization approach, e.g., the algorithm SKATER (Spatial' K'luster Analysis by Tree Edge Removal) proposed by Assunção *et al.*, 2006, satisfies our requirements, i.e., find clusters taking into account the spatial contiguity of the areas (ATS) analyzed.

In particular, this approach summarized the spatial units by their centroids, modeled as a node in an undirected graph. The spatial constraint is based on the spatial neighborhood structure inside the undirected graph defined by geographical adjacency, i.e., the spatial areas share at least one boundary or vertex. The spatial clusters are then defined as connected subgraphs that minimize the within-cluster heterogeneity, computed by associating a weight to each edge that connects the nodes (i.e., ATS) i and j. These weights are based on dissimilarity measures between locations i and j concerning their attribute vectors x_i and x_j (i.e., variables describing the supply and demand of childcare services). In general, the simple Euclidean squared distance between them is considered a measure of dissimilarity. The within-clusters similarity is described as the Euclidean squared distance between the location attributes in cluster k and the cluster means of these attributes.

The SKATER algorithm uses the minimum spanning tree approach to reduce the graph complexity. In brief, the minimum spanning tree is a spanning tree (i.e., sub-graph of an undirected connected graph, which includes all the vertices of the graph with a possible minimum number of edges) in which the sum of the weight of the edges is as minimum as possible. After constructing the minimum spanning tree, the SKATER algorithm prunes the tree for the desired number of clusters to minimize the within-clusters variability.

The paper is organized as follows. Section 2 describes the data used in the analysis. Section 3 shows the results coming from the SKATER algorithm, while Section 4 summarizes the conclusions and further research.

2. Data description

In this work, we use two data sources to capture the complex and multidimensional structure of demand and supply of childcare services within Italy. We analyze the 2019 data on childcare services from the ISTAT survey "Survey of childcare and early childhood supplementary services" (ISTAT, 2019b) that describes the supply side of the welfare. Relevant cultural and social features that can describe the demand side, such as the fertility rate, the presence of family support (e.g., grandparents, babysitters), and characteristics of the families (e.g., number of members, educational level of the parents), were included in the analysis using data from the permanent census (ISTAT, 2019a).

Table 1 provides a brief overview of the variables used in this analysis. The first two (i.e., coverage rate and per capita expenditure rate) represent the supply side, while the remaining variables describe the demand side. The latter variables were chosen to define possible solutions that can be broadly considered as alternative solutions to child care (e.g., babysitter, grandparent, extended family) and outline the socioeconomic characteristics of the population in each area. Therefore, several correlated variables are analyzed. However,

the multicollinearity does not impact our results since the variables enter in the clustering approach as weights of the graph described as the sum of squared Euclidean distances. So if two variables are strongly correlated, they simply enter twice into the weights's definition, but for all ATS.

Table 1. List of the variables analyzed. The first two variables describe the supply side of childcare services, while the remaining outline the demand side.

Variable	Description
Coverage	N. of day-care places/n. of resident children between 0-2 years old.
Per capita expenditure rate	(Computed on the 2 years old resident population).
Female employment rate	N. of resident females working/n. of the total residents.
Female house rate	N. of not working resident females at home/n. of resident children from 0 to 2
Commuter rate	N. of commuters for work outside the municipality/n. of workers
Male educational qualification rate	N. of male persons with a degree higher than high school degree/n. of male persons over 20 years of age
Female educational qualification rate	N. of female persons with a degree higher than high school degree/n. of female persons over 20 years of age
Foreign rate	N. of foreign residents/number of the total residents
Grandparent rate	N. of resident retired persons/n. of resident children from 0 to 2
Babysitter rate	N. of not working resident females (studying) from 15 to 25 years old/n. of resident children from 0 to 2
N. household's members	N. of members in the household
Fertility rate	N. of resident children with age 0/ n. of residents between 15 and 49 ages.
Social and Material Vulnerability Index (IVSM)	Composed of 7 different indicators. For more details, please refer to ISTAT (2020).

Source: ISTAT (2019a), ISTAT (2019b).

The analysis is developed at the administrative unit (ATS) level, i.e., inter-municipal aggregations that handle social programming and, therefore, often intersect with the

scheduling of childhood education services. From a theoretical and political perspective, the choice to focus on this micro-scale level to highlight the within-regional variability of childcare services could be motivated by historical factors, both related to demand dynamics (e.g., declining birth rates, women's participation in employment) and from local and regional government initiatives in the absence of effective regional governance. Therefore, the analysis of the childcare service is conducted at an intermediate scale between the provincial and municipal dimensions, i.e., the ATS level. In this way, a more appropriate picture of childcare services' demand and supply characteristics can be obtained, taking into account their distribution in areas larger than the municipalities and, simultaneously, bringing out specific differences between specific provinces.

We focus on a single year (i.e., 2019) to have available and up-to-date data and on one region (i.e., the Veneto region), having seven provinces, 21 ATS, and 563 municipalities.

3. Results

In this section, we show the results from the SKATER algorithm imposing the number of clusters equal to 6 corresponding to the first plateau of the Elbow function. Figure 1 shows the geographical positions of these clusters, while Figure 2 describes the distribution of the variables (scaled) defined in Table 1 by corresponding boxplots for each cluster.

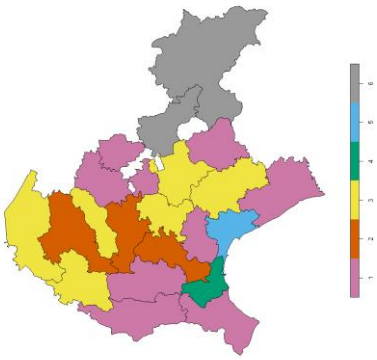


Figure 1. Geographical representation of the 6 clusters created by the SKATER algorithm.

The first cluster (pink-colored) is composed of ATS from several areas of the region. Looking at the corresponding boxplots in Figure 2, we can identify this cluster as territories with a smaller number of children, i.e., low fertility rate (average equals -0.506), few members in the household (average equals -0.098), but with some alternatives to childcare services, i.e., quite high grandparents (average equals 0.367) and babysitter rates (average equals 0.448). It can be considered as the “average cluster.”

The second cluster (orange-colored) comprises ATS within the municipalities of Verona, Vicenza, and Padova. It is a more advanced welfare model: the coverage is high (average equals 1.019), as well as the female employment participation (average equals 0.806) and the educational level of both parents (average equals 1.604 for males and 1.614 for females). This cluster includes three main municipalities of the Veneto region characterized by universities, work opportunities, and road connections.

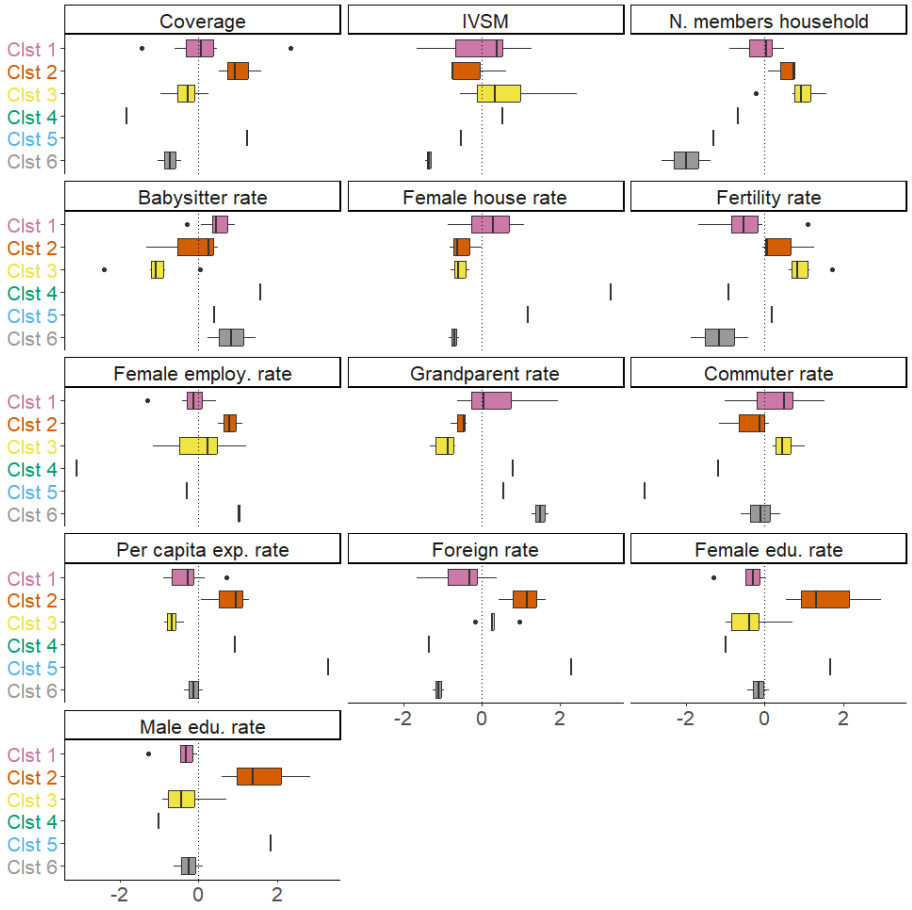


Figure 2. Boxplot for each variable (scaled) of Table 1 and cluster calculated by the SKATER algorithm.

The third cluster (yellow-colored) is the most critical regarding socio-economic situations. The territory, which includes ATS from the peripherical areas of Verona, Vicenza, Treviso, and Padova, is characterized by low supply and relatively high demand. We note a high fertility rate (average equals 0.97), a high number of members in the household (average equals 0.861), and a low coverage (average equals -0.328) and per capita expenditure rate

(average equals -0.675). The socio-economic situation of these territories is more problematic: the women work (average equals 0.059), the fertility rate is high, but the level of education is pretty low (average equals -0.355), there are no alternatives to childcare services, and there is social and material deprivation. Here, stakeholders must focus on adding new *kindergarten* places or alternatives that can reach the lower middle section of society.

The fourth cluster (green-colored) comprises the Venetian territories of Chioggia, Cona, and Cavarzere, where childcare welfare must be improved. We can observe low coverage (average equals -1.841), female employment rate (average equals -3.101), and male/female educational level (average equals -1.009 for males and -1.007 for females). There may be a need for more job opportunities for women in Chioggia since the actual predominant sector is traditionally male (fishing, agriculture, glass, and wood processing).

The fifth cluster (blue-colored) includes the Venezia municipality, with a very low commuter rate (average equals -3.054) and a significantly high per capita expenditure rate (average equals 3.302). The population has a high educational level (average equals 1.85 for males and 1.664 for females) but few job opportunities for women (average equals -0.301). However, educational services for children and possible alternatives are, for well-known choices of policies, available in this area.

The sixth cluster (grey-colored) comprises the mountain area: the province of Belluno, characterized by an aging population. We can note here a high grandparents rate (average equals 1.49) as well as a low number of members in the household (average equals -1.996).

4. Conclusions

We wanted to show how gathering data from different sources and applying suitable statistical models can help in reading the territories and favor tailored policies on relevant issues for the populations. In this paper, we analyze and emphasize the variability of childcare services' demand and supply characteristics in the Veneto region in Italy, thanks to the SKATER regionalization approach. This method helps deal with the data's intrinsic spatial structure and the complex high-dimensional structure of the variables analyzed. In general, Veneto is a region with, on average, good coverage of childcare services, with a wide range of daycare centers and preschools. However, the coverage of preschool services in Veneto varies by area and municipality. The regionalization approach used highlights, in fact, several sub-regional models within the Veneto region. We noted how childcare services are more focused around the main municipalities, particularly those characterized by various work opportunities, both for females and males, and by the presence of universities. The suburban areas are instead affected by a lack of childcare services as well as job opportunities for women. When new welfare policies are applied to these territories,

policymakers must consider these differential socioeconomic characteristics: taking into account only the supply distribution on its own could lead to inefficient policies. Thus, for the first time, to our knowledge, a spatial clustering method has been applied to synthesize multidimensional data to help policymakers understand where to develop new efficient policies to level out inequalities in the supply and demand for *kindergartens* within Italy. The analysis can be extended to other regions or states by choosing accurately the variables that describe the complex system of children's services. This is just an example of how, using current and existing data, much information to produce the "intelligence of data" to inform public policies.

References

Assunção, R. M., Neves, M. C., Câmara, G., & Da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7), 797-811.

Borghorst, M., Mulalic, I., & Van Ommeren, J. (2021). Commuting children and the gender wage gap. *Technical report, Tinbergen Institute Discussion paper*, (No. TI 2021-089/VIII)

Brilli, Y., Del Boca, D., & Pronzato, C. D. (2016). Does child care availability play a role in maternal employment and children's development? Evidence from Italy. *Review of Economics of the Household*, 14, 27-51.

Del Boca, D. (2002). The effect of child care and part time opportunities on participation and fertility decisions in Italy. *Journal of population economics*, 15, 549-573.

Dipartimento per le politiche della famiglia, ISTAT, Università Ca' Foscari, and MIPA (2020). Nidi e servizi educativi per la prima infanzia, stato dell'arte, criticità e sviluppi del sistema educativo integrato 0-6, Roma.

ISTAT (2019a). Censimento permanente della popolazione e delle abitazioni. https://www.istat.it/it/censimenti-permanenti/popolazione-e-abitazioni.

ISTAT (2019b). Indagine su nidi e servizi integrativi per la prima infanzia. https://www.istat.it/it/archivio/267291.

ISTAT (2020). Le misure della vulnerabilità: un'applicazione a diversi ambiti territoriali. *Letture statistiche – Metodi*, Roma.

Landivar, L. C., Ruppanner, L., & Scarborough, W. J. (2021). Are States Created Equal? Moving to a State With More Expensive Childcare Reduces Mothers' Odds of Employment. *Demography*, 58(2), 451-470.

Plantenga, J., Remery, C., & Camilleri-Cassar, F. (2009). The provision of childcare services: A comparative review of 30 European countries. Luxembourg: European Commission's Expert Group on Gender and Employment Issues (EGGE).

Measuring energy poverty in Spain with the new EU expenditure-based indicators

Judit Mendoza Aguilar[1], Francisco J. Ramos-Real[1], Alfredo J. Ramírez-Díaz[2]
[1]Departament of Economics, Accounting and Finance, Universidad de La Laguna, Spain, [2]Energías Sin Fronteras, Canary Islands section, Spain.

Abstract

This paper analyzes energy poverty in Spain between 2016 and 2021, using the new European primary indicators that relate household income to their energy expenditure, called expenditure-based indicators. The objective of the study is to determine the characteristics of the households most vulnerable to energy poverty in Spain, that is, with a greater probability of incurring in this situation. The determinants that influence energy poverty are identified through machine learning models: logistic regression using bootstraping and random forest using repeated cross-validation. The problem addressed is key in the current economic and regulatory context of the energy transition, and it is essential to provide tools to measure its impact and analyze the causes.

Keywords: *energy poverty; logistic regression, primary indicators; machine learning; random forest.*

Exploring the impact of websites on hospital services in Puerto Rico: analyzing opportunities and challenges in Healthcare Administration through Internet and Social Media integration

Dharma Vazquez Torres[1], Michael Concepción-Santana[1]
[1] Department of Health Services Administration, University of Puerto Rico, Puerto Rico

Abstract

This study explores how websites affect hospital services and Puerto Rico Health System website integration possibilities and issues. Technology has improved hospital patient care, engagement, and efficiency (Korda & Itani, 2011). This study analyzes Puerto Rico's hospital websites content and patient involvement. "About the Hospital" and "Contact Us" were the most popular website components in a 68-hospital descriptive survey. "Healthcare Research" and "Education and Training" were the least publicized on social media, with 30% of hospitals. The study shows that good communication and technology improve patient care and engagement. Private, non-profit, and state hospitals websites were examined for content, patient education, institution type, clinical services, facilities and amenities, conditions and treatments, news and events, job possibilities, Facebook, Twitter, and YouTube linkages, and patient and visitor information. These criteria were evaluated as binary variables if present in all sample hospitals. This study will contribute to digital technology in healthcare literature and offer Puerto Rican and worldwide hospital administration and healthcare practitioners useful advice.

Keywords: *hospital websites, patient engagement, healthcare communication*

Korda, H., & Itani, Z. (2013). Harnessing social media for health promotion and behavior change. Health Promotion Practice, 14(1), 15–23. https://doi.org/10.1177/1524839911405850

Lupton, D. (2014). The commodification of patient opinion: The digital patient experience economy in the age of big data. Sociology of Health & Illness, 36(6), 856–869. https://doi.org/10.1111/1467-9566.12109

Ventola C. L. (2014). Social media and health care professionals: benefits, risks, and best practices. P & T : a peer-reviewed journal for formulary management, 39(7), 491– 520.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16440

AI in the newsroom: A data quality assessment framework for employing machine learning in journalistic workflows

Laurence Dierickx[1], Carl-Gustav Lindén[1], Andreas L Opdahl[1], Sohail Ahmed Khan[1], Diana Carolina Guerrero Rojas[1]
[1]Department of Information Science and Media Studies, University of Bergen, Norway

Abstract

AI-driven journalism refers to various methods and tools for gathering, verifying, producing, and distributing news information. Their potential is to extend human capabilities and create new forms of augmented journalism. Although scholars agreed on the necessity to embed journalistic values in these systems to make AI-driven systems accountable, less attention was paid to data quality, while the results' accuracy and efficiency depend on high-quality data. However, defining data quality remains complex as it is a multidimensional and highly domain-dependent concept. Assessing data quality in AI-driven journalism requires a broader and interdisciplinary approach, considering journalists as end-users. It means meeting the challenges of data quality in machine learning and the ethical challenges of using machine learning in journalism. These considerations ground a conceptual data quality assessment framework that aims to support the collection and pre-processing stages in machine learning. It aims to strengthen data literacy in journalism by emphasizing limitations and possible biases related to data and making a bridge between journalism studies and scientific disciplines that should be viewed through the lenses of their complementarity.

Keywords: *data quality assessment, journalism, ethics, machine learning, artificial intelligence*

This research was funded by EU CEF grant number 2394203.

1. Introduction

AI-driven journalism refers to various methods and tools for news gathering, verification, production, and distribution (Thurman et al., 2019). They aim to support professional practices to help speed up time-consuming tasks, publish automated content, identify trends, or provide insights into large numeric or textual datasets. Hence, their potential is to extend human capabilities and augment journalism practices (Lindén, 2018). Although AI-driven systems are often considered opaque and not bias-free (Guidotti et al., 2019), they depend on high-quality data to avoid inaccurate analytics and unreliable decisions (Gupta et al., 2021). Explaining how data is collected, organised, cleaned, annotated, and processed participates in establishing a relationship of trust between the journalist as the end-user and the tool. It implies understanding the challenges of data quality that appear upstream and downstream of a machine learning process (Gudivada et al., 2017).

The "garbage in, garbage out" principle also applies in journalism, whereas quality information requires quality data to ensure the accuracy and reliability of the news (e.g., Anderson, 2018; Diakopoulos, 2020; Dierickx, 2017; Dörr & Hollbuchner, 2017; Lowrey et al., 2019). However, less attention was paid to this critical aspect. The conceptual framework presented in this paper intends to fill this gap, considering that assessing data quality is context and use dependent (Tayi & Ballou, 1998; Boydens & Van Hooland, 2011).

2. Theoretical backdrops

Data quality encompasses several complementary dimensions referring to a set of attributes in which dimensions – such as accuracy, completeness, and consistency – were refined over time. However, research agreed that data quality refers to data that adapts to the uses of data consumers, especially in terms of accuracy, relevance, and understandability (Wang & Strong, 1996). The emergence of big data brought new challenges, such as believability, verifiability, and the reputation of the data (Batini et al., 2015). The level of trustability of the data was also underlined, as various data sources challenge their interoperability and the contexts where data are used (e.g., Cai & Zhu, 2015; Liu et al., 2016; Saha & Srivastava, 2014). Big data quality issues are also related to incomplete, inaccurate, inconsistent, or ambiguous structured and unstructured data (Eberendu, 2016).

Approaching data quality in machine learning includes all these considerations but also encompasses several particularities insofar as the quality of the results is influenced by the data provided as input to the system (Gudivada et al., 2017; Gupta et al., 2021). Also, research emphasised that models trained on incomplete or biased datasets can produce discriminatory outputs and interfere with the accuracy of the tasks (Miceli et al., 2022; Shin et al., 2022).

Data quality issues are likely to appear since the data acquisition stage: data availability does not equal data quality (Elouataoui et al., 2022), especially when working with open data, user-generated data, or data coming from multiple sources (Hair & Sarstedt, 2021). Data pre-processing involves addressing classical data quality issues, such as missing data, duplicates, strongly correlated variables, abnormal or inconsistent values, normalization, and standardization (Polyzotis et al., 2018; Foidl & Felderer, 2019; Elouataoui et al., 2022).

Training datasets, which refer to the process of adapting the model to the data, are needed to evaluate the suitability of the data for machine learning tasks –in terms of efficiency, accuracy and complexity (Gupta et al., 2021). In this context, the validation process aims to ensure that data does not contain errors that can propagate into the model. These errors will likely be introduced during the collection, aggregation or annotation stage (Polyzotis et al., 2018; Gupta et al., 2021). However, it is practically impossible to achieve it exhaustively, even though evaluating the risk of poor data quality is possible. At the same time, there is a lack of discussion on methods to define the level of validation in each step of a machine learning process (Foidl & Felderer, 2019). Furthermore, corpus annotations for supervised tasks are problematic because they are inherently error-prone, either if they rely on automation or crowdsourcing (Gupta et al., 2021).

Because the relationship between users and AI systems lies on trust (Rai, 2020), data quality should follow three fundamental principles: prevention, detection, and correction to ensure the trustworthiness and reliability of machine learning applications (Ehrlinger et al., 2019). A good understanding of the data provides correct analyses and reliable decisions (Gupta et al., 2021). It should also reflect the knowledge of the domain experts. Furthermore, selecting or creating a dataset for an AI-driven system involves human decisions beyond technical aspects, thus requiring empirical considerations (Miceli et al., 2021).

Ethical journalism practices join these concerns. They refer to the rules, routines and institutionalised procedures to produce knowledge (Ekström, 2002). Although ethical journalism is a question of practice, providing truthful information is not dissociable from the news's credibility (or believability) (van Dalen, 2019). Hence, ethical principles of journalism can be summarised according to the main principle of respecting the truth with accuracy and objectivity (Ward, 2018). The development of data-driven practices focused specifically on the data source's reliability, accuracy, the right to extract and use the data, and the right to privacy (Craig et al., 2017). At the same time, transparency has become a motto, viewed as an instrument to increase credibility and trust toward audiences (Koliska, 2022).

In AI-driven journalism, the ethical challenges of transparency concern the data, the algorithms at work, and the outcomes (Dörr & Hollbuchner, 2017). Nonetheless, transparency is not always easy to implement in journalism, where practitioners often lack data and algorithm literacy to grasp how algorithms work (Porlezza & Eberwein, 2022), no

more than it is easy to implement in deep learning models where even their creators need to learn how they operate because of the multiplicity of their parameters (Burkart & Huber, 2021). While a recognised need exists to blend AI-driven systems with journalistic values to fit professional practices (Broussard et al., 2019; Gutierrez Lopez, 2022), it should start with the data. If they are biased or contain errors, the system will likely reproduce these biases and errors (Hansen et al., 2019). Considering that accuracy and reliability are two prerequisites of ethical journalism practices, trusting the system is also about trusting the data it relies on.

3. Building the conceptual assessment framework

Data quality assessment is critical and gives rise to operations that aim to improve the overall data quality by identifying erroneous data elements and understanding their impact on the processes at work (Cichy & Rass, 2019). From an end-user perspective, assessing data quality indicators refer to their fitting to human needs or user requirements through the aggregation of different information on data quality (Cappiello et al., 2004). The assessment framework we have developed in the context of AI-driven journalism is a part of this data quality assessment tradition. It is based on the learnings from the scientific literature (e.g., Batini et al., 2009; Cichy & Rass, 2019; Fox et al., 1994; Pipino et al., 2002; Shanks, 1999) and on the core ethical principles in journalism acknowledged by professionals.

The ethical principle of telling the truth relates to respecting facts. It refers to the syntactic and semantics levels and the dimensions of the data's accuracy, consistency, correctness, and understandability. It requires the application domain knowledge to deal, for instance, with incorrect values or duplicates. Because objectivity is a disputed concept in journalism due to its intrinsic subjective nature, we privileged the one of fairness related to the elements that guarantee to report honestly, avoiding bias or unbalanced information. It concerns the context of producing, validating, disseminating, and using the data for a journalistic purpose. Hence, it is connected to the pragmatic level and relates to the dimensions of timeliness, completeness, accessibility, objectivity, relevance, and usability. Transparency refers to the trustability of information, but it is not the only constituent of trust. The broader concept of trust can be thus understood through the social semiotic level. It encompasses the dimensions of credibility, reliability, and verifiability.

The assessment framework encompasses formal and empirical indicators, inducing that the overall assessment includes a human perspective. Its application can be objective or subjective (Pipino et al., 2002) to detect data quality issues and challenges likely to appear upstream of the processes, either generally or more granularly. It can be applied to the data collection and pre-processing stages from which the training, the test, and the validation datasets are derived for developing machine learning systems in a journalistic context.

Table 1. Data quality assessment framework.

Ethical	Semiotic	Dimension	Verification
Truth	**Syntactic**	Accuracy	- Level of interoperability, standardisation - Measure of erroneous data (ratio accurate values/total values) - Uniqueness (duplicate entries and redundancies) - Encoding problems and information overload
		Consistency	- Well-defined data structure (percentage of data with consistent format and values) - Homogeneity vs heterogeneity (format, structure, values) when data come from multiple sources - Unambiguous and explicit labelling
	Semantic		
		Correctness	- Identifying abnormal values - Identifying the causes of NULL values - Evaluation of the spelling coherence - Data documented/compliant with metadata
		Understandability	- The extent to which data are comprehensible (feedback from the end-user)
Fairness	**Pragmatic**	Timeliness	- Currentness (percentage of updated data)
		Completeness	- Appropriate amount of data (ratio missing values/total values, ratio NULL values/total values)
		Accessibility	- Right to use the data (terms of use) - Level of retrievability of the data
		Objectivity	- Unbiased data (size and representativity)
			- Identification of human bias (annotation incl.)
		Relevance	- The extent to which the data are relevant for the purpose (feedback from the end-user)
			- Newsworthiness (journalistic added values and expected impact, feedback from the end-user)
			- Data scarcity (measurement of the fraction of data containing relevant information)
		Usability	- Fitness for use (to assess globally through the formal and empirical indicators of the frameworks, = making sense of AI in a journalistic context)
			- How automation structures and presents the data
Trust	**Social**	Reliability	- Authenticity (source) - Authority (source, annotators) - Reputation (source, annotators)
		Credibility	- Degree of the believability of the data source
			- Degree of the believability of the data
			- Degree of the believability of the annotation process and of the annotators
		Verifiability	- Verification of the source and the data
			- Verification of the annotation process

This framework was applied to a sample of datasets used for automated fact-checking. While the syntactic level did not have particular issues, on the semantic level, a cross-domain approach and a strong language dependency challenged the understandability and correctness of the datasets. The pragmatic level appeared problematic due to NULL values, no attached licence, and no mention of the last update. The dimension of completeness was more difficult to assess because of the content's domain and language dependency. A lack of harmonization

in the classification was also detected, from "true" to "false", "half true", "contradiction", or "unrelated". On the social level, datasets collected from Wikipedia raised questions about their reliability and credibility, due to the participative nature of this platform.

4. Conclusion

Acknowledging that the relationship between journalists and AI-driven systems is built on trust, the data that feed these systems must also be trusted. However, the definition of "good" data in journalism remains challenging due to the multidimensionality of the concept of quality. This concept is intrinsically related to the expertise of a given application domain and to the understanding of how data are collected, validated, and disseminated. It should also be considered through its relevance to be used in a journalistic context and the overall purpose of the AI-driven system. Also, approaching data quality through normative lenses consists of a practical solution to address the recognized need for embedding journalistic values and ethical principles in AI-driven systems. Hence, the conceptual assessment framework presented in this communication was designed as an adaptive and flexible tool that can be used in various forms that AI-driven journalism tools can take. It shows that data quality issues are far from trivial, as the quality of the data at every stage of the process will directly influence machine learning outcomes. Nevertheless, the main limitation of this framework is that it is only applicable for common machine learning tasks because the provenience and nature of the vast amounts of data used in the most complex systems remain mostly uncertain.

References

Anderson, C. W. (2018). *Apostles of certainty: Data journalism and the politics of doubt*. Oxford University Press.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52.

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From data quality to big data quality. *Journal of database management*, *26*(1), 60–82.

Boydens, I., & van Hooland, S. (2011). Hermeneutics applied to the quality of empirical databases. *The Journal of Documentation; Devoted to the Recording, Organization and Dissemination of Specialized Knowledge*, *67*(2), 279–289.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245-317.

Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., & Chuan, C.-H. (2019). Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly*, *96*(3), 673–695.

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the Big Data era. *Data Science Journal*, *14*(0), 2.

Cappiello, C., Francalanci, C., & Pernici, B. (2004). Data quality assessment from the user's perspective. In *Proceedings of the 2004 international workshop on Information quality in information systems* (pp. 68–73).

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 24634–24648.

Craig, D., Ketterer, S., & Yousuf, M. (2017). To post or not to post: Online discussion of gun permit mapping and the development of ethical standards in data journalism. *Journalism & Mass Communication Quarterly*, *94*(1), 168–188.

Diakopoulos, N. (2019). *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.

Dierickx, L. (2017). News bot for the newsroom: How building data quality indicators can support journalistic projects relying on real-time open data. In *Global Investigative Journalism Conference 2017, Academic Track*.

Dörr, K. N., & Hollnbuchner, K. (2017). Ethical challenges of algorithmic journalism. *Digital Journalism*, *5*(4), 404–419.

Eberendu, A. C.. (2016). Unstructured Data: an overview of the data of big data. *International Journal of Computer Trends and Technology*, *38*(1), 46–50.

Ehrlinger, L., Haunschmid, V., Palazzini, D., & Lettner, C. (2019). A DaQL to monitor data quality in machine learning applications. In *Lecture Notes in Computer Science* (pp. 227–237). Springer.

Ekström, M. (2002). Epistemologies of TV journalism: A theoretical framework. *Journalism*, *3*(3), 259-282.

Elouataoui, W., Alaoui, I. E., & Gahi, Y. (2022). Data quality in the era of big data: A global review. In *Big Data Intelligence for Smart Applications* (pp. 1–25). Springer.

Foidl, H., & Felderer, M. (2019). Risk-based data validation in machine learning-based software systems. *Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation*.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, *30*(1), 9–19.

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, *10*(1), 1–20.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42.

Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., Mehta, S., Guttula, S., Afzal, S., Sharma Mittal, R., & Munigala, V. (2021). Data quality for machine learning tasks. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Gutierrez Lopez, M., Porlezza, C., Cooper, G., Makri, S., MacFarlane, A., & Missaoui, S. (2022). A question of design: Strategies for embedding AI-driven tools into journalistic work routines. *Digital Journalism*, 1–20.

Hair, J. F., Jr, & Sarstedt, M. (2021). Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. *The Journal of Marketing Theory and Practice*, *29*(1), 65–77.

Hansen, M., Roca-Sales, M., Keegan, J. M., & King, G. (2017). *Artificial intelligence: Practice and implications for journalism*. Columbia University.

Karlsen, J., & Stavelin, E. (2014). Computational journalism in Norwegian newsrooms. *Journalism Practice*, *8*(1), 34–48.

Keller, S., Korkmaz, G., Orr, M., Schroeder, A., & Shipp, S. (2017). The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application*, *4*(1), 85–108.

Koliska, M. (2022). Trust and journalistic transparency online. *Journalism Studies*, *23*(12), 1488–1509.

Lindén, C.-G. (2018). What Makes a Reporter Human? A Research Agenda for Augmented Journalism. *Questions de communication*, (37), 337–351.

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing: Official Publication of the International Society for Photogrammetry and Remote Sensing (ISPRS)*, *115*, 134–142.

Lowrey, W., Broussard, R., & Sherrill, L. A. (2019). Data journalism and black-boxed data sets. *Newspaper Research Journal*, *40*(1), 69–82.

Miceli, M., Posada, J., & Yang, T. (2021). Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction*, *6*, 1–14.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218.

Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2018). Data lifecycle challenges in production machine learning: A survey. *SIGMOD Record*, *47*(2), 17–28.

Porlezza, Colin, & Eberwein, T. (2022). Uncharted territory: Datafication as a challenge for journalism ethics. In *Media and Change Management* (pp. 343–361). Springer.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141.

Saha, B., & Srivastava, D. (2014). Data quality: The other face of big data. *2014 IEEE 30th International Conference on Data Engineering*.

Shanks, G. (1999). Semiotic approach to understanding representation in information systems. In *Proceedings of the Information Systems Foundations Workshop, Ontology, Semiotics and Practice*.

Shin, D., Hameleers, M., Park, Y. J., Kim, J. N., Trielli, D., & Diakopoulos, N. (2022). Countering algorithmic bias and disinformation and effectively harnessing the power of AI in media. *Journalism & Mass Communication Quarterly*, *99*(4), 887–907.

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, *41*(2), 54–57.

Thurman, N., Lewis, S. C., & Kunert, J. (2019). Algorithms, automation, and news. *Digital Journalism*, *7*(8), 980–992.

van Dalen, A. (2019). Journalism, Trust, and Credibility. In *The Handbook of Journalism Studies* (pp. 356–371). Routledge.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, *12*(4), 5-33.

Ward, S. J. A. (2018). Reconstructing journalism ethics: Disrupt, invent, collaborate. *Media & Jornalismo*, *18*(32), 9–17.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16434

Study of the relationship between competitiveness and digital footprint indicators in Valencian wineries

Leonardo Castro[1], Ana Debón[1], Josep Domenech[2]
[1]Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Spain,
[2]Department of Economics, Universitat Politècnica de València, Spain.

Abstract

The digital footprint of the Spanish wine sector is a valuable resource for predicting real-time indicators, enabling companies to anticipate their competitors and devise effective digital transformation strategies using emerging technologies. With advances in computation and web-scraping techniques, it is now possible to approximate competitiveness indicators using real-time information from company websites. Given this context, the general objective of this work is to analyze the relationship between the digital footprint and competitiveness of Valencian wine companies. To this end, it is proposed to use financial variables obtained from the Sistema de Análisis de Balances Ibéricos (SABI) and indicators extracted from the companies' websites. Unsupervised learning techniques will be implemented to find groups or clusters of companies based on their economic performance. Subsequently, digital footprint indicators will be used to create a supervised learning model to predict the above classification of companies based solely on digital footprint indicators to identify the most significant indicators for predicting competitiveness.

Keywords: *Digital footprint; web scraping; competitiveness; supervised and unsupervised learning; wineries.*

1. Introduction

The Spanish wine sector holds great value and relevance in the country's economy, society, and culture. According to data from the Spanish Wine Federation, Spain has approximately 13% of the total number of vineyards in the world, making it the third-largest producer and exporter of wine in terms of volume. The sector supports around 427,700 direct and indirect jobs and represents 2.2% of the Gross Value Added in Spain. Additionally, the sector has been characterized by heavy investment in innovation and development, with between 170 and 180 million euros per year invested in R&D activities over the last five years.

Due to intense competition in the market and the trend of companies toward digital transformation, the wine sector has realized the need to implement digital strategies for communication, customer acquisition, commercialization, and marketing of its products. Since digital transformation is critical to increasing business competitiveness, wine companies and wineries have increased their online presence through online sales platforms, corporate websites, and social networks, generating a significant digital footprint. This digital footprint can be detected and measured in various ways to monitor economic characteristics with real-time indicators. These indicators are valuable to private and public organizations as they enable them to anticipate competitors and identify strategies to implement within the digital transformation framework and emerging technologies (Blazquez et al. 2018).

In this context, this research aims to analyze the relationship between the digital footprint and competitiveness in Valencian wineries using multivariate techniques.

2. Materials and Methods

2.1. Data

The data matrix comprises 115 observations and 56 variables from wine companies in the Valencian Community. All companies are identified by name and their official website.

The variables used in the analysis are financial and digital footprint indicators. The financial variables were obtained from the SABI (Iberian Balance Sheet Analysis System) database, using the same variables as Rodriguez (2022). On the other hand, the digital footprint indicators used were those suggested by Blazquez and Domenech (2018), plus some others regarding company online activities on social networks.

2.2. Multivariate Analysis

Statistical analyses were conducted using the R environment for statistical computing (R Core Team, 2023). Clustering was used to identify groups of wine companies based on their financial variables, differentiating the sample of companies according to their

competitiveness characteristics. Subsequently, footprint indicators were used to create a Generalized Linear Model (GLM, McCullagh, 2019) to predict the above classification of companies using only digital footprint indicators.

Then, Receiver Operating Characteristics (ROC) graphs to evaluate the concordance between the models and actual data. Calculating the area under the ROC curve of the classifier (AUC) is a standard method to reduce ROC performance to a single scalar value representing expected performance. Therefore, as the area under the ROC curve (AUC) increased, the classifier power also increased

3. Results

After pre-processing our data, we tested clustering algorithms with different numbers of clusters (k) using the silhouette coefficient as a guide. Therefore, we decided to choose k=2 for our final analysis. We applied both K-means and fuzzy clustering algorithms to our dataset. Figure 1 shows the results obtained by applying fuzzy clustering, which resulted in two clusters with distinct financial characteristics.

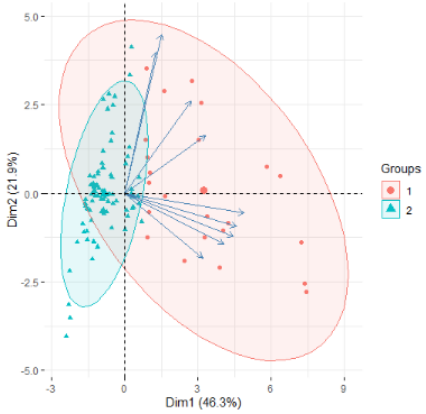


Figure 1. Results of fuzzy clustering. Source: own elaboration.

Twenty-three companies were assigned to Cluster 1, whereas 88 were assigned to Cluster 2. Cluster 1 comprises companies with higher mean values for all their competitiveness indicators; these metrics are associated with competitive performance, indicating that companies in this cluster have superior competitive performance compared to those in Cluster 2.

The logistic regression model used variables such as the age of the domain, the availability of an English version of the website, the number of Instagram posts, the number of Instagram followers, and the presence of specific keywords (such as export, efficiency,

performance, novelty, competitiveness, differentiation, LinkedIn, immaterial, brands, value, network, positioning, country, and change), using broad match based on stemming.

The model's estimated results showed a high predictive capacity, with an accuracy of 0.81 and an AUC of 0.80.

4. Conclusions

The main objective of this study was to explore the association between digital footprint indicators and competitiveness among Valencian wineries. Through exploration of the sample of companies using financial variables obtained from SABI, two clusters of companies with distinct competitiveness characteristics were identified. The first cluster consisted of companies exhibiting high levels of competitive performance and innovation capabilities, with more employees, capital, trademarks, and economic performance. On the other hand, the second cluster was composed of smaller companies with fewer employees and lower economic returns compared to the first cluster, indicating a more local market orientation.

Then, a logistic regression model was constructed, including the following variables: domain age, the English version of the website, the number of Instagram posts, the number of Instagram followers, and the presence of specific keywords. These keywords included export, efficiency, performance, novelty, competitiveness, differentiation, LinkedIn, immaterial, brands, value, network, positioning, country, and change. The model demonstrated high predictive capacity, with an accuracy of 0.81 and an AUC of 0.80.

Finally, it can be concluded that a significant relationship exists between the digital footprint indicators and a company's competitiveness, which can be used to differentiate between the previously identified clusters.

References

Blazquez, D., Domenech, J. (2018). Web data mining for monitoring business export orientation. *Technological and Economic Development of Economy*, 24(2), 406-428.

McCullagh, P. (2019). *Generalized linear models*. 2nd Edition. Routledge.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rodríguez Zamora, EE. (2022). Influencia de la actividad digital en la competitividad de las bodegas valencianas. Universitat Politècnica de València. http://hdl.handle.net/10251/186563.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16508

Solo Consumption – A machine learning approach

Aikaterini Manthiou[1], Van Ha Luong[2], Phil Klaus[3]
[1]Marketing, Neoma Business School, France, [2]Marketing, IESEG School of Management, France, [3]Marketing, International University of Monaco, Monaco.

Abstract

This study aims at conceptualizing the solo tourism consumption journey. We use a semi-supervised machine learning approach and analyze more than 27,000 tweets. The seed sets extraction, seed and topic confidence and model fit evaluations will provide us with the dimension of solo tourism conceptualization.The results will reveal how consumers perceive solo tourism consumption. This study provides scholars and managers with an evidence-based solo consumption conceptualization, as well as with a marketing, psychological, and operation tool to manage the solo consumer segment.

Keywords: *Solo tourism; semi-supervised Latent Dirichlet Allocation (LDA); machine learning approach; tweets*

1. Introduction

Despite the emergence of solo consumption during the pandeminc, it is currently an under-researched area lacking a comprehensive and systematic examination (Leith, 2020; Otegui-Carles, Araújo-Vila, & Fraiz-Brea, 2022). Research has to date analyzed solo tourism in an inconsistent way, which is why it warrants an updated and thorough investigation (Yang, Nimri, & Lai 2022). This paper aims to provide a holistic and all-inclusive view of the solo tourism consumption byt using twitter data (Otegui-Carles, Araújo-Vila, & Fraiz-Brea, 2022).

Our study contributes in multiple ways to both consumer research and management. First, the study helps identify the key elements of a concise solo tourism construct. Second, this study allows us to assess the existing solo consumption research's disintegrated state and develop a more comprehensive illustration of it. Third, this research uses a semi-supervised machine learning approach to analyze user-generated contents and validate the solo consumption framework. Fourth, we provide managers and marketers with a valuable tool for comprehending the triggers of solo consumption and the desired experiences.

2. Literature review

Solo tourism is a state of being alone during a trip (Yang et al., 2022). Specifically, solo tourism is the activity of tourists traveling to destinations alone and for various reasons (Jonas, 2022). Solitary consumers are therefore those who choose to travel on their own (Bianchi, 2022) or individuals dining alone at restaurants (Choi et al., 2022). They are described on the basis of different factors, such as their personal needs, desires, motivations, preferences, and travel behavior (Leith, 2020).

Due to the COVID-19 pandemic, tourists perceive solo tourism as a secure travel option, and a recovery action for the tourism and hospitality industry (Jonas, 2022). Owing to its complexity, there is no unified agreement on or an all-inclusive understanding of the solitary consumption in existing research. Consequently, a consensual, comprehensive conceptualization of solo tourism consumption is required (Otegui-Carles, Araújo-Vila, & Fraiz-Brea, 2022).

3. Methodology

Since the purpose of our study is to scrutinize the meaning of solo tourism through solo tourists' viewpoints, we used the Twitter API (application programming interface) and the hashtags #solotravel and #solotourism to collect all related tweets in English published between August 2019 and August 2022. In total, we collected 43,290 unprocessed tweets. We will employ a semi-supervised LDA and the keyATM package to investigate and

empirically validate the solo tourism dimensions. Two steps will be employed. In step 1 we extract seed sets (Watanabe & Zhou, 2022) and step 2 we work on the Seeded-LDA model training (Benoit et al., 2018). Moreover, based on seed sets extraction, seed and topic confidence and model fit evaluations will provide us with exact dimension of solo tourism conceptualization. Appendix 1 depitcs the algorithm.

4. Conclusion

Our study will advance both, solo consumption theory and managerial practice. Our work will deliver a more detailed and overall conceptualization of the solo consumption concept. In particular, this research will provide a valuable tool for managers and marketers to comprehend the triggers of solo consumption and the desired experiences that tourists who engage in these activities are looking for. Highlighting the solo tourism experience's different stages allows managers to determine if, and if so how, they can integrate the solo tourism segment into their marketing strategy. We therefore provide managers with guidance on how to design and market solo offers and activities effectively. Last, we will use an innovative method, a semisupervised machine learing approach to reach the goals of the study.

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30), 774.

Bianchi, C. (2022). Antecedents of tourists' solo travel intentions. Tourism Review, 77(3) 780-795.

Choi, S. H., Cho, M., Yang, E. C. L., & Tabari, S. (2022). Emotional congruence among solo diners. International Journal of Hospitality Management, 101, 103-108.

Jonas, L. C. (2022). Solo tourism: A great excuse to practice social distancing. African Journal of Hospitality, Tourism and Leisure, 11(SE1), 556-564.

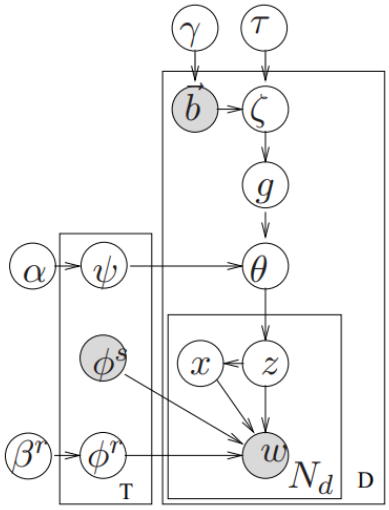
Leith, C. (2020). Tourism trends: Lifestyle developments and the links to solo tourism. Journal of Tourism Futures, 6(3), 251-255.

Otegui-Carles, A., Araújo-Vila, N., & Fraiz-Brea, J. A. (2022). Solo travel research and its gender perspective: A critical bibliometric review. Tourism and Hospitality, 3(3), 733-751.

Watanabe, K., & Zhou, Y. (2022). Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. Social Science Computer Review, 40(2), 346-366.

Yang, E. C. L., Nimri, R., & Lai, M. Y. (2022). Uncovering the critical drivers of solo holiday attitudes and intentions. Tourism Management Perspectives, 41, 100913.

Appendix 1. Seeded LDA model and algorithm



1. For each k=1$\cdots$T,
   (a) Choose regular topic $\phi_k^r \sim \text{Dir}(\beta_r)$.
   (b) Choose *seed* topic $\phi_k^s \sim \text{Dir}(\beta_s)$.
   (c) Choose $\pi_k \sim \text{Beta}(1, 1)$.
2. For each seed set $s = 1 \cdots$ S,
   (a) Choose group-topic distribution $\psi_s \sim \text{Dir}(\alpha)$.
3. For each document d,
   (a) Choose a binary vector $\vec{b}$ of length S.
   (b) Choose a document-group distribution $\zeta^d \sim \text{Dir}(\tau\vec{b})$.
   (c) Choose a group variable $g \sim \text{Mult}(\zeta^d)$.
   (d) Choose $\theta_d \sim \text{Dir}(\psi_g)$. // of length T
   (e) For each token $i = 1 \cdots N_d$:
      i. Select a topic $z_i \sim \text{Mult}(\theta_d)$.
      ii. Select an indicator $x_i \sim \text{Bern}(\pi_{z_i})$.
      iii. if x_i is 0
         - Select a word $w_i \sim \text{Mult}(\phi_{z_i}^r)$.
      iv. if x_i is 1
         - Select a word $w_i \sim \text{Mult}(\phi_{z_i}^s)$.

Exploring the use of machine learning and explainability in Marketing Mix Modeling

Slava Kisilevich[1], Markus Hermann[1]
[1]Lidl Analytics, Lidl International, Germany

Abstract

Marketing Mix Modeling (MMM) employs statistical techniques, typically linear regressions, to assess the impact of advertising expenditure on sales. Despite advancements in statistics and machine learning, the field of MMM has remained relatively unchanging due to a few reasons: (1) its primary focus on practical business applications, (2) the proprietary nature of MMM solutions by specialized companies, and (3) the difficulty in interpreting complex models beyond linear regressions for business purposes.

Recently, there has been increased emphasis on the interpretability of complex machine learning models. To address this, model explainers such as SHAP have been introduced, enabling the application of non-linear machine learning algorithms in the realm of MMM. This provides a solution to the various issues associated with traditional MMM methods, including variable interactions, non-linear relationships, and interpretability.

This presentation outlines a method for incorporating machine learning algorithms with explainability techniques in the context of MMM in the retail industry

Keywords: *MMM; Marketing Mix Modelling; Machine Learning; MMM Explainability; SHAP.*

A simple and efficient kNN variant with embedded feature selection

Almudena Moreno-Ribera[1], Aida Calviño[1]
[1]Department of Statistics and Data Science, Complutense University of Madrid, Spain.

Abstract

Predictive modeling aims at providing estimates of an unknown variable, the target, from a set of known ones, the input. The k Nearest Neighbors (kNN) is one of the best-known predictive algorithms due to its simplicity and well behavior. However, this class of models has some drawbacks, such as the non-robustness to the existence of irrelevant input features or the need to transform qualitative variables into dummies, with the corresponding loss of information for ordinal ones. In this work, a kNN regression variant, easily adaptable for classification purposes, is suggested. The proposal allows dealing with all types of input variables while embedding feature selection in a simple and efficient manner, reducing the tuning phase. More precisely, making use of the weighted Gower distance, we develop a powerful tool to cope with these inconveniences by implementing different weighting schemes. The proposed method is applied to a collection of 20 data sets, different in size, data type and the distribution of the target variable. Moreover, the results are compared with previously proposed kNN variants, showing its supremacy, particularly when the weighting scheme is based on non-linear association measures and in datasets that contain at least one ordinal input variable.

Keywords: *Gower distance; weighting scheme; ordinal variables; Machine Learning; predictive modeling; regression.*

Optimization techniques for Kernel Logistic Regression on large-scale datasets: A comparative study

José Ángel Martín-Baos[1], Ricardo García-Ródenas[1], Luis Rodriguez-Benitez[2]
[1]Department of Mathematics, University of Castilla-La Mancha, Spain, [2]Department of Information and System Technologies, University of Castilla-La Mancha, Spain.

Abstract

In recent years, machine learning techniques have been increasingly applied to modelling the decision-making processes of individuals. One technique that has shown good results in the literature for modelling complex behaviours is the Kernel Logistic Regression (KLR). However, standard KLR implementations have a time complexity of $\mathcal{O}(n^3)$, which is not feasible for large datasets. To overcome this limitation, one of the purposed alternatives is to approximate the kernel matrix using the Nyström method. The aim of this work is to evaluate the Nyström KLR model on large-scale datasets and to study, at the experimental level, which of the optimisation techniques that allow training this model is the most efficient. As results, the authors show that the Nyström method efficiently computes the objective function and its gradient, enabling the training of KLR models with up to 10^5 parameters. Then, it is evaluated the performance of several optimisation methods, including gradient descend, Momentum, Adam, and L-BFGS-B. It can be concluded that L-BFGS-B is the most efficient method for training the Nyström KLR model. However, given enough computational time and proper hyperparameter tuning, the Adam method can also yield good results.

Keywords: *Random Utility Model; Machine Learning; Kernel Logistic Regression; Nyström method; Numerical Optimisation.*

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16416

Use of machine learning techniques in non-probabilistic samples

Jorge Rueda[1], Beatriz Cobo[2], Luis Castro[2]
[1]Department of Statistics and Operations Research, University of Granada, Spain, [2]Department of Quantitative Methods for Economics and Business, University of Granada, Spain.

Abstract

Non-probabilistic surveys are increasingly used because they are easy and cheap to carry out. Even official statistical agencies are starting to use this type of surveys in their research, due to the difficulty and the amount of resources needed to carry out probabilistic surveys, which are currently the best option due to their reliability. When non-probabilistic surveys are used, the classical estimation methods cannot be used since the initial conditions for carrying them out are not met, so over the years new estimation techniques have been emerging in this type of sampling. Some of the most relevant estimation techniques currently being used are those related to machine learning techniques.

In this work we focus on the estimation technique for non-probabilistic samples statistical matching, which can be enhanced and improved if we complement it with a machine learning technique known as XGBoost. We are going to study a variable of interest extracted from a real non-probabilistic survey carried out during the COVID-19 pandemic, and check if by applying such estimations we obtain better results than without applying this type of techniques.

Keywords: *Machine learning; non-probabilistic sampling; statistical matching; XGBoost.*

1. Introduction

The major strength of probability sampling is that the probability selection mechanism permits the development of statistical theory to examine the properties of sample estimators. The weakness of all nonprobability methods is that no such theoretical development is possible; as a consequence, nonprobability samples can be assessed only by subjective valuation (Kalton, 1983). Over the years the development of non-probabilistic surveys has boomed and many techniques have been developed to calculate reliable estimates from non-probabilistic survey data.

Many advanced in artificial intelligence models, such as deep learning techniques, have shown remarkable accuracy in prediction. Artificial intelligence models perform poorly when dealing with relatively small data sets, while machine learning models have good predictive performance on smaller data sets. However, a single machine learning approach often leads to overfitting and difficulty handling the large number of imbalanced data sets that occur in real world problems. To make up for the shortcomings of a single machine learning method, the conjoint learning technique based on the GBDT (Gradient Boost Decision Tree) algorithm was developed and has gradually become the mainstream approach in the field of learning research automatic. eXtreme Gradient Boosting (XGBoost) is a highly efficient booster set learning model originated from the decision tree model, which uses the tree classifier for better prediction results and higher operational efficiency.

This technique has been used in many settings, for example Li and Yao (2018) classify gene mutations using machine learning models, XGBoost and SVM, in the hope of improving gene mutation classification performance. In terms of performance of the two qualifying models, XGBoost outperformed SVM. From the confounding metrics, it could be seen that XGBoost had better predictive ability, especially for those with enough classes featured. Liu et al. (2021) used a mortality prediction model using the XGBoost decision tree model for patients with acute kidney injury in the intensive care unit, and compared its performance with that of three other machine learning models, logistic regression (LR), support vector machines (SVM), and random forest (RF) being XGBoost the best performing algorithm in this study. Castro-Martin et al. (2021) test the potential of the XGBoost algorithm in the most important estimation methods that integrate data from a probability survey and a non-probability survey. At the same time, a comparison is made of the effectiveness of these methods for the elimination of biases. The results show that the proposed estimators based on gradient increasing frameworks can improve the representativeness of the survey with respect to other classical prediction methods. The proposed methodology is also used to analyze a real sample from a non-probabilistic survey on the social effects of COVID-19. Cui et al. (2022) created an accurate prediction model using machine learning techniques, such as logistic regression, XGBoosting machine,

random forest, neural network, gradient boosting machine, and decision tree, to predict 3-month mortality specifically among lung cancer patients with bone metastases according to readily available clinical data. Today, people tend to use credit cards for their payment efficiency, but credit cards also provide a new opportunity for fraud. Companies and researchers have been trying to come up with a method to tell if a transaction is fraudulent. Cai and He (2022) propose a hybrid model based on the combination of TabNet and XGBoost. A dataset provided by IEEE-CIS is used in this investigation, which contains many records of transactions and whether they are fraudulent.

Our work focuses on the combination of data obtained through probabilistic and non-probabilistic surveys with the aim of obtaining more reliable estimates through XGBoost. As a non-probabilistic survey, we will base ourselves on the survey carried out by Pérez et al. (2020) and as a probabilistic survey the CIS Barometer of May 2020.

2. Methodology

Let U be the finite population of interest of size N, s_v a non-probabilistic (or volunteer) sample of size n_v, from which we measure a vector of auxiliary variables $x = (x_1, \dots, x_p)$ and the variable of interest *y* that we want to know about the population U. Normally the results we obtain from this kind of samples present different types of biases, especially the one known as selection bias, which appears if there is a significant difference between the individuals in our sample and those not sampled. To correct this type of bias there are several techniques, which depend on the type of auxiliary information available (Rueda et al., 2020). If we have a reference probability sample s_r, of which we only know the same vector of auxiliary variables as in s_v, we can apply the technique known as statistical matching, based on superpopulation models.

2.1. Statistical Matching (SM)

Also known as Mass Imputation, it was developed by Rivers (2007). It is based on modeling the relationship between the variable of interest and the vector of auxiliary variables, using the non-probabilistic sample s_v to predict the values of the variable of interest in the probabilistic sample s_r, since they are unknown. Assuming that the population of interest U is a realization of a superpopulation model *m*:

$$y_i = m(x_i) + e_i, \;\; i = 1, \dots, N$$

Where $m(x_i) = E_m[y_i|x_i]$ y $e \sim N(0, \sigma)$. That is, we can model the relationship between the variable of interest and the auxiliary variables using some model (which we will call SM). From such a model we estimate the prediction of the values of *y* in the probability sample s_r, using the values of the auxiliary variables in that sample, of the form:

$$\hat{y}_i = E_{SM}[y_i|x_i, 1_i], \;\; i \in s_r$$

1_i will have a value equal to one if the i-th individual belongs to the probability sample s_r, and will be zero when it does not belong to this sample. Depending on the model we use, we will have different expressions of $\hat{y}_i$. Once we obtain the prediction of our variable of interest *y*, we can construct the estimator of our choice in the form (case of the estimator of the population total):

$$\hat{Y}_{SM} = \sum_{i \in s_r} \hat{y}_i \cdot w_{ri}$$

Being w_{ri} the design weight for the i-th element of the reference sample. We see that in this technique the most important step is to predict the variable of interest *y*, to perform this step we can use machine learning techniques that produce a prediction as accurate as possible. In our work we will use the technique known as XGBoost, which is producing excellent results both in the prediction of variables and in the estimation of inclusion probabilities for non-probabilistic samples.

2.2. XGBoost Estimator

In our case we will use the XGBoost technique to obtain the predicted values of the response variable for the probabilistic sample s_r. This machine learning technique works as a group of decision trees, which establish branches (different paths) as a function of x_i until a final value $\hat{y}_i$ is obtained (Chen and Guestrin, 2016). The expression of $\hat{y}$ using XGBoost is:

$$\hat{y}_i^{XG} = \phi(x_i) = \sum_{k=1}^{K} f_k\,(x_i), \;\; f_k \in F$$

where K is the number of decision trees, $F = \{f(x) = \omega_{q(x)}\}$ with $q: \mathbb{R}^p \rightarrow T$ representing the structure of each tree, and ω_i is the score of the i-th final node. Finally we obtain the predicted value $\hat{y}_i^{XG}$ by summing the scores of each tree, which are designed to minimise the following objective function:

$$L(\phi) = \sum_i l\,(\hat{y}_i^{XG}, y_i) + \sum_k \Omega\,(f_k)$$

where l is a function that measures the error in the estimates, and which must be differentiable and convex (i.e. difference squared). To regularise this function, there is a $\Omega(f)$ that penalises trees with too many final nodes *T* and exaggeratedly high scores ω, of the form:

$$\Omega(f) = \gamma T + \frac{\lambda ||\omega||^2}{2}$$

being γ and λ hyperparameters that directly influence the regularisation of the function. This regularisation serves to control the so-called overfitting, which appears when the machine learning model has a behaviour specific to the type of data we train it with, producing bad results when the input data are different to those we have used to train the model (Hawkins, 2004). Because of this it is very important what values these hyperparameters have, that can be taken arbitrarily or by hyperparameter optimization (i.e. by cross validation). Finally $L(\phi)$ is minimised with the gradient tree boosting method, developed by Friedman (2001). This allows us to converge to the minimum value of a function through an iterative process (gradient descent), training the models by giving more importance to the data for which previous models have failed (boosting). To improve its performance, XGBoost also implements other techniques such as shrinkage, to limit the influence of each individual tree, among others (Chen and Guestrin, 2016).

Once we estimate the values of the variable of interest for the individuals of the probability sample s_r by XGBoost $\hat{y}_i^{XG}$, we obtain that the estimator of the population total using statistical matching is:

$$\hat{Y}_{SM}^{XG} = \sum_{i \in s_r} \hat{y}_i^{XG} \cdot w_{ri}$$

3. Application

Combining statistical matching with XGBoost as the chosen machine learning method is a relatively costly process which, in addition, has to be repeated for each variable of interest. In this case, we have chosen the following variable from the survey conducted by Pérez et al. (2022) during the Spanish lockdown caused by the COVID-19 pandemic: "Would you be willing to continue teleworking after the lockdown?". We could then evaluate the interest of the population in working remotely now that, even though it is not mandatory anymore, it has emerged as an interesting option.

The percentage of individuals responding affirmatively considering only the non-probabilistic sample in a naive way would be 26.2%. However, it is preferable to consider possible biases caused by the snowball methodology used during the distribution of the online survey. For this reason, we also consider the CIS Barometer of May 2020. The variables in common between our non-probabilistic sample and the auxiliary probabilistic sample are the following: state, province, urban density, sex, age, education level, employment status, last electoral vote, intended electoral vote and confidence in the government during the pandemic.

Once the bias reduction process is completed, we find out that the percentage of individuals who would not mind continuing to telework is actually 33.1% instead of the initial 26.2%. Therefore, we observe a significant increase from the initial impression before considering a more advanced analysis.

4. Conclusions

In this work, we have considered a method combining statistical concepts and advanced machine learning techniques in order to improve the reliability of the estimations for a variable of interest. We have also observed, via a real application, how relevant applying said method can be for the final conclusions obtained.

When a strict methodology is not considered for carrying out a survey, it is important to consider these kinds of methods in order to avoid possible biased results.

References

Cai Q. & He J. (2022). Credit Payment Fraud detection model based on TabNet and Xgboot. 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, pp. 823-826, doi: 10.1109/ICCECE54139.2022.9712842.

Castro-Martín, L., Rueda, M.M., Ferri-García, R., & Hernando-Tamayo, C. (2021). On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures. *Mathematics* 9, 23: 2991. https://doi.org/10.3390/math9232991.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Cui, Y., Shi, X., Wang, S., Qin, Y., Wang, B., Che, X., & Lei, M. (2022). Machine learning approaches for prediction of early death among lung cancer patients with bone metastases using routine clinical characteristics: An analysis of 19,887 patients. *Frontiers in Public Health*, 10. https://doi.org/10.3389/fpubh.2022.1019168.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Hawkins, D. M. (2004). The problem of overfitting. Journal of chemical information and computer sciences, 44(1), 1-12.

Kalton, G. (1983). *Introduction to Survey Sampling.* Newbury Park, CA: Sage Publications.

Li, G., & Yao, B. (2018). Classification of Genetic Mutations for Cancer Treatment with Machine Learning Approaches. *International Journal Of Design, Analysis And Tools For Intergrated Circuits And Systems*, 7(1), 63-67.

Liu, J., Wu, J., Liu, S., Li, M., Hu, K. & Li, K. (2021) Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. *Plos One* 16(2): e0246306. https://doi.org/10.1371/journal.pone.0246306.

Peréz, V., Aybar, C. & Pavía, J.M. (2022). Dataset of the COVID-19 lockdown survey conducted by GIPEyOP in Spain. *Data in Brief*, 40, https://doi.org/10.1016/j.dib.2021.107700.

Rivers, D. (2007, August). Sampling for web surveys. In Joint Statistical Meetings (Vol. 4).

Rueda, M., Ferri-García, R., & Castro, L. (2020). The R package Non-ProbEst for estimation in non-probability surveys. The R Journal, 12(1), 406-418.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16452

Optimizing floor price in Real Time Bidding

David Gávez[1], Víctor Dugo[2]
[1]Departamento de Estadística e Investigación Operativa , Universidad de Sevilla, Spain,
[2]Departmento de Análisis Económico y Economía Política, Universidad de Sevilla, Spain

Abstract

Between 150 and 200 words briefly specifying the aims of the work, the main results obtained, and the conclusions drawn

AdNetwork companies are very much a part of today's new digital marketing methods. This paper aims to develop an algorithm that solves the problems of AdNetwork companies in setting optimal floor prices. Establishing the optimal starting price for the bid is equivalent to setting the price that maximises revenue, which is optimal for the publisher and the AdNetwork company. In this market, that price will balance two opposite scenarios: a high floor price could lead to some impressions unsold, while a low floor price could be insufficient to reach profit margins. The contribution is twofold. First, this paper extends the problem of optimal price flor in real time bidding auctions for advertising in current scenarios where a DSP (Demand Side Platform) acts as a filter and only one bid is received by the AdNetwork and thus, the price paid corresponds to the reserve price. and, moreover, it is implemented in reality with a pseudo-algorithm (not provided for commercial reasons). It allowed to be implemented in a real case scenario for three publishers, obtaining an average increase of revenue of 127%.

Keywords: *AdNetwork; DSP; optimal floor price; regret.*

Editorial Universitat Politècnica de València

1. Introduction

Digital marketing has become companies' primary advertising tool and has changed how firms communicate with customers (Chaffey & Ellis-Chadwick, 2019). Within this field, some companies operate through what is known as an 'AdNetwork'. An AdNetwork can essentially be defined as a network that acts as a link between advertisers and publishers (sites or websites that sell their advertising space), thus generating income for publishers and placing advertisers' resources in the most appropriate places for the advertising content (D'Annunzio & Russo, 2020; Tahaei & Vaniea, 2021). An AdNetwork has a dual revenue source: on the one hand, it generates income from the advertising campaigns it sells or manages for advertisers. On the other hand, it administers the publishers' advertising space and takes a share of the revenue generated by the advertisements shown (impressions) for offering the technology, posting and linking. The revenue share determines this part of the price.

The management of publisher inventory is done by purchasing advertising impressions in this inventory. The AdNetwork will then seek advertisers who will pay the price ("revenue" from now on) to advertise in that inventory by purchasing those impressions. It does this through blinded second-price real-time auctions (Myerson, 1981) in which the auction to acquire those inventory impressions is won by the advertiser who bids the most but pays the price of the second highest bid, which is the final "revenue". These auctions have a floor or reserve price built into them, which is the minimum price the publisher will receive for ad impressions in its inventory. Impressions are the number of times an ad appears in that inventory, and the price is usually per thousand impressions.

To this "revenue" is applied the share agreement so that the publisher will receive the agreed percentage of the revenue, and if this percentage results in a value below the floor price, the floor price is paid, and the company (AdNetwork) reduces its profit margin. Thus, if there is a revenue share commitment of 70/30 (70% corresponds to the publisher while the remaining 30% is the AdNetwork company's profit for its services), this means that Adnetwork auctions the impressions from the publisher's inventory at a price where at least 70% of that price corresponds to the floor price set by the publisher. Thus, for example, if a publisher sets a floor price of 7 euros, Adnetwork will put it up for auction at 10 euros in order to maintain its revenue share of 70/30 and will not accept bids below that price, which is the price that allows it to meet the publisher's floor price and maintain its profit.

Therefore, establishing the optimal starting price for the bid is equivalent to setting the one that maximizes revenue, which is optimal for the publisher and the "AdNetwork" company. In this market, that price will balance two opposite scenarios: a high floor price could lead to some impressions unsold, while a low floor price could be insufficient to reach profit margins. The development of the article is as follows. In Section 2, The proposed approach

and different scenarios are presented. In Section 3 the problem is mathematically formalized and developed. Section 4 presents some particular cases due to different auction platform practices. Finally, Section 5 provides some actual results and conclusions.

2. Justification of the approach and scenarios.

2.1. Justification of the approach

The reliability of the algorithm is established according to the concept of regret. This concept answers the question: at the end of T iterations of the algorithm, where we have all the information of what happened, what would happen if it had simply applied the same decision rule **h** at each iteration? One could calculate the loss of this fixed hypothesis by adding the personal loss of the T iterations. If this value is less than the loss incurred by doing different actions, the decision maker is incurring *regret*, which is the difference between these two losses because we could have chosen a single action each iteration and obtained better results than we did. Typically, the regret is calculated with respect to the optimal strategy known after the period is over, i.e. ex-post, so, at each iteration, it is necessary to choose the action that is understood to minimize the regret (although this is not known until the end of the period).

Let us assume that each loss for each iteration of T is between 0 and 1, so the total loss at the end of the period will be between 0 and T. For a hypothesis and a loss function, if the algorithm guarantees that for all possible states, the regret is $\breve{O}(T)$, it means that as T tends to infinity, the average regret per iteration tends to zero (since $\breve{O}(\cdot)$ is the rate at which regret converges to zero in each iteration), and there are T iterations.

In other words, if we design an algorithm and implement it T times, it incurs a loss after that period. The goal is to avoid the situation where seen in retrospect (a posteriori), the algorithm has incurred a lower loss with a constant rule of decision. Thus, the regret of an algorithm is the difference between the loss of the algorithm and the loss from using the constant alternative.

Translating this to our problem, the algorithm sets a floor price. After the period, the revenue obtained is compared with the one that would have resulted from applying the optimal price calculated a posteriori based on the actual data and the loss is obtained for having set that price and not the optimal one in that period. This comparison is the regret. The algorithm has been set to have a regret $\breve{O}(T^{1/2})$, which is a good result and means that when T tends to infinity, the regret converges to zero in $T^{1/2}$, which implies that it needs a smaller number of steps for the regret to converge.

2.2. Scenarios

2.2.1. Initial scenario

In an initial scenario, AdNetwork companies faced repeated auctions as a seller, each with a different number of bidders. Each bidder will submit a bid, which will be an unknown and random value to other bidders' eyes. The "revenue" is the second highest bid price, and the aim is to maximize the profit by considering only the "revenues" from previously recorded bids. A fair starting price (reserve or floor price) must be set to do this. The algorithm established a mechanism to find the optimal starting price that maximizes "revenue" by accessing only two pieces of information: the final "revenues" of the previous auctions and the number of bidders that finally participate in each auction (both pieces of information are obtained as outputs given by the digital ad auction platforms). After the algorithm's development, the game's initial rules changed.

2.2.2. Current scenario

In the current scenario, the different bidders have been replaced, and bids are now placed through Google's DSP, and Google only sends the winning bid. Under these new conditions, there is only access to one bid above the floor price; therefore, whatever the bid's value is, it ends up paying the floor price with the consequent loss of profitability. This is why it is necessary to create a new algorithm to establish an optimal price within the new market rules.

3. Formal problem statement and development

3.1. Mathematical formulation

Without loss of generality, the algorithm has been designed for prices between 0 and 1, with 0 being the minimum price and 1 the maximum possible price. This only requires rescaling the actual prices according to these limits. The algorithm launches a reserve price, observes what happened and does the necessary calculations. Based on these calculations, it chooses a new price to maximize the expected revenue and launches it, repeating the process. The algorithm was first proposed in (Cesa-Bianchi et al., 2015) as follows.

The firm conducts an auction to sell an item. In the initial scenario, after the auction, it collects $m \geq 2$ bids: $B_1, B_2, \ldots, B_m$, which are observations of m independent and identically distributed (i.i.d.) random variables. As indicated, prices will be between zero and one, so $B_i \in [0,1]$, $i = 1, \ldots, m$. These random variables have a common distribution function (since they are i.i.d) F, arbitrary and unknown, which will show the probability that a bid is below a certain value. We will denote $B^{(1)}, B^{(2)}, \ldots, B^{(m)}$ by the statistical order of the bids such that $B^{(1)} \geq B^{(2)} \geq \cdots \geq B^{(m)}$.

The algorithm will set a starting price (reserve price hereafter) $p \in [0,1]$ for the auction, and after the auction is conducted, it observes the revenue $R(p)$, which will depend on the chosen reserve price, and the values of the bids, i.e. $R(p) = R(p; B_1, B_2, \ldots, B_m)$ which is defined as:

$$R(p) = \begin{cases} B^{(2)} & if \quad p \leq B^{(2)} \\ p & if \quad B^{(2)} < p \leq B^{(1)} \\ 0 & if \quad p > B^{(1)} \end{cases}$$

That is, if bids are received below p (or no bids are received), the item is not sold, and the "revenue" is zero, and if bids are received above that price, the item is sold to the bidder who bids $B^{(1)}$ at the price of the second highest bid, i.e. $B^{(2)}$, this $B^{(2)}$ being the revenue. If the item is sold, the middle condition guarantees that the revenue will always have the reserve price as the minimum price. However, this information is only obtained a posteriori. When the algorithm launches a price p, it does so, expecting a revenue $\mu(p) = E[R(p)]$, which is the expected revenue when the algorithm uses the price p. With the appropriate mathematical manipulations, the expected revenue can be rewritten as:

$$\mu(p) = \int_p^1 x \, dF_2(x) + pP(B^{(2)} < p \leq B^{(1)}) = E[B^{(2)}] + \int_0^p F_2(t) \, dt - p(F(p))^m$$

where $F_2(x)$ denotes the probability that $B^{(2)}$, is less than or equal to x (distribution function of $B^{(2)}$ and $(F(p))^m$ is the joint distribution function of $B^{(1)}, B^{(2)}, \ldots, B^{(m)}$ and the price that maximizes it: $$p^* = \arg \max_{p \in [0,1]} \mu(p)$$

This expected revenue will depend on the bid distribution function, i.e. F. The algorithm allows for cases where more than one bid is received as long as this number constitutes a percentage less than the parameter α to be defined later.

However, in the current scenario where the platform registers only one bid (winning bid in a DSP where there are several bidders) and therefore the advertiser always pays the floor price, and there is only access to the value of the winning bids, which are always (or 1-α% of the total number of times) the only one above the floor price and to the number of bidders in the DSP, if P_F is the advertiser floor price agreed with the publisher, the revenue function is:

$$R(p) = \begin{cases} P_F & if \quad B^{(1)} \leq P_F \\ 0 & if \quad P_F > B^{(1)} \end{cases}$$

And the scenario differs from the one analyzed in (Cesa-Bianchi, Gentile and Mansour, 2015). It is different in that information is only available for the highest bid and not for the second bid, as the floor price is always paid when only one bid is received from the Demand Side Platform. That is, everything must be inferred from the distribution function of the highest bid F_1. Let F_1 denote the distribution function of $B^{(1)}$. This function will indicate the probability that the winning bid is less than or equal to a certain value. Thus, $F_1(x)$ indicates the probability that the value of the highest bid, i.e. $B^{(1)}$, is less than or equal to x, while $F_2(x)$, as indicated, indicates the probability that $B^{(2)}$, is less than or equal to x.

At this point, it is necessary to establish a lemma whose demonstration has been developed but can also be mathematically intuited. The algorithm is going to raise the floor price in order to receive it as revenue constantly. It will raise it according to the values of the first bids and the revenue it obtains from the net increase in impressions that it stops obtaining when advertisers are unwilling to pay these higher floor prices. In this circumstance, if the algorithm can only access the winning bids on the floor price and the number of bidders in the DSP, it is possible to establish the revenue function and optimal price as:

$$p^* = \arg \max_{p \in [0,1]} \mu(p) = \arg \max_{p \in [0,1]} E\left[B^{(2)}\right] + \int_0^p F_2(t)dt - p(F(p))^m$$

$$\equiv \arg \max_{p \in [0,1]} E\left[B^{(1)}\right] + \int_0^p F_1(t)dt - p(F(p))^m = \arg \max_{p \in [0,1]} \mu^{(1)}(p)$$

As the function $(F(p))^m$ is not available, an approximation is computed from what is available, which is the distribution $F(1)$.

$F_1(p) = \beta((F(p))^m) = m((F(p))^m)\frac{m-1}{m} - (m-1)((F(p))^m)$ for $m \geq 2$

So: $$\mu^{(1)}(p) = E[B^{(1)}] + \int_0^p F_1(t)\, dt - p\beta^{-1}(F_1(p))$$

If $m = 1$, then the joint distribution function $(F(p))^m$ corresponds to the observed distribution, i.e. $F_1(p)$.

From here, the designed algorithm works in each auction as follows. In auction t, it will set the price p_t, and will have a revenue after the auction of $R_t(p_t) = R(p_t; B_{t,1}, B_{t,2}, \ldots, B_{t,m})$ which is a function of the random variables $B_{t,1}, B_{t,2}, \ldots, B_{t,m}$ in the auction or time t. The price given by the algorithm will depend on the previously observed $B^{(1)}$ and the floor price, and therefore on the past bids, as it learns and updates from them.

Thus, given a sequence of reserve prices $p_1, p_2, \ldots, p_T$ set by the algorithm, the cumulative regret up to T will be given by:

$$\Sigma_1^T(\mu(p^*) - \mu(p_t))$$

Therefore, the regret (the reliability or goodness of an algorithm) will be a random variable since it depends on p_t, which will depend on the previous revenues that will depend on $B_1, B_2, \ldots, B_m$.

It is important to note that we do not have access to the actual distribution of $B^{(1)}$, which is unknown, but to its empirical distribution function, which allows us to calculate the equivalent revenue $\mu^{(1)}(p)$ that provides the same maximizer as the expected revenue $\mu(p)$.

The algorithm works in stages. In each stage, the algorithm is run a certain number of times with the same reserve price. This is necessary to obtain the empirical distribution function of $B^{(1)}$, i.e. F_1. In principle, it is assumed that the algorithm will run a total number of T times.

- Stage 1 will contemplate T_1 auctions (implementations of the algorithm), and therefore the price it will use will be $p_t = \hat{p}_1;\ t = 1, \ldots, T_1$.
- Stage 2 will contemplate T_2 auctions (implementations of the algorithm), and therefore the price it will use will be $p_t = \hat{p}_2;\ t = T_1 + 1, \ldots, T_1 + T_2$.
- And so on.

In this way, the algorithm will produce reserve prices of $0 = \hat{p}_1 \leq \hat{p}_2 \leq \ldots \leq 1$. They are set from an interval built according to a signification level α, choosing the price from this interval that minimizes risk subject to constraints related to the distribution function. The total number of stages is denoted as S (stages). It is shown mathematically that for the algorithm to have the agreed regret, each stage must have a number of implementations or auctions $T_i = T^{1-2^{-i}}$. From here, the number of stages, or at least their upper limit, can be determined so that $S \leq \lceil 2\log_2 \log_2 T \rceil$, i.e. it shall be set to the smallest integer not less than $2\log_2 \log_2 T$. The total cumulative regret of the algorithm shall be:

$$\Sigma_1^S(\mu(p^*) - \mu(p_i))T_i$$

4. Special cases

4.1 Treatment of the algorithm when the number of bids is not known

In this case, the number of bids is not known. However, a limited number of different floor prices can be set for different advertisers. In this way, each advertiser who wants to advertise sees a different price set, depending on the algorithm that predicts how much they are willing to pay. According to empiric approaches (Seljan et al., 2014; Ballesteros et al., 2015), it could be assumed that the number of bidders follows a discrete normal distribution:

$$H(m) = P(M = m) = \frac{e^{\frac{-1}{2\sigma^2}(m-\mu_m)^2}}{\Sigma_{m_i} e^{\frac{-1}{2\sigma^2}(m_i-\mu_m)^2}}, m_i = -\infty, \dots, -1, 0, +1, \dots, +\infty$$

Thus, the expected revenue will be: $\mu(p) = E_M E[R^M(p)] = E_M E\left[B_M^{(2)}\right] + \int_0^p E_M[F_2, M](t)dt - pE_M[F^M](p)$

Where $E_M[F^M](p)$ can be estimated from the support function $T(x) = \Sigma_{m=2}^{\infty} H(m)x^m$ and its auxiliary function $A(x) = T(x)(1-x)T'(x)$ with the appropriate mathematical steps (Cesa-Bianchi et al., 2015).

5. Conclusions

This paper extends the problem of optimal price flor in real time bidding auctions for advertising in current scenarios where a DSP acts as a filter and only one bid is received by the AdNetwork and thus, the price paid corresponds to the reserve price. It also materialized the case in which the number of bidders is unknown using the normal discrete distribution. After an evaluation period carried out by an Andalusian digital marketing agency (Creafi), the revenue after using the algorithm described in section 4 increased by 127% compared to the revenue obtained if the default price agreed with the publisher had been used.

References

Ballesteros-Pérez, P., González-Cruz, M. C., Fuentes-Bargues, J. L., & Skitmore, M. (2015). Analysis of the distribution of the number of bidders in construction contract auctions. Construction management and economics, 33(9), 752-770.

Cesa-Bianchi, N., Gentile, C. and Mansour, Y. (2015). Regret Minimization for Reserve Prices in Second-Price Auctions, IEEE Transactions on Information Theory, 61(1), pp. 549–564.

Chaffey, D., & Ellis-Chadwick, F. (2019). Digital marketing. Pearson UK.

D'Annunzio, A., & Russo, A. (2020). Ad networks and consumer tracking. Management Science, 66(11), 5040-5058.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research*, *6*(1), 58-73.

Tahaei, M., & Vaniea, K. (2021). "Developers Are Responsible": What Ad Networks Tell Developers About Privacy. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-11).

Yuan, S., Wang, J., Chen, B., Mason, P., & Seljan, S. (2014, August). An empirical study of reserve price optimisation in real-time bidding. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1897-1906).

Hotel price forecasting using time series. An exploratory research

Esther Chávez-Miranda[1], Sergio Toral[2], M. Rocío Martínez-Torres[3]
[1]Department Financial Economy and Operations Management, Universidad de Sevilla, Spain, [2]Department of Electronic Engineering, Universidad de Sevilla, Spain, [3]Department of Business Administration and Marketing, Universidad de Sevilla, Spain.

Abstract

This paper proposes the use of time-series-based forecasting methods to identify the main predictor variables of prices in hotels located in the city of Barcelona. However, in contrast to previous works, the research focusses on online prices, i.e., the prices set by hotel companies' revenue management algorithms, rather than purchase prices. For the training of the time series, a dataset of hotel prices offered on Booking.com with a horizon of zero days in advance has been used. In addition to the price series itself, a set of exogenous variables has been included to improve the predictive capacity of the model. As a result, the relative importance of the lags of the endogenous variables and of the exogenous variables, as well as the prediction error, have been obtained. The lag is the main variable in the determination of the forecast and, more specifically, those referring to one day-, one week-, and one month-lags.

Keywords: *Hotel price; Time series; autoregressive models; Revenue management.*

This publication is part of the project TED2021-130406B-I00, funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU"/PRTR, and the authors would also like to thank the company BeonX, S.L. for their collaboration

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16505

Some empirical observations on price patterns in online stores

Álvaro Gómez-Losada[1], Néstor Duch-Brown[2]
[1]Department of Quantitative Methods, Universidad Loyola Andalucía, Seville, Spain, [2]Digital Economy Unit, Joint Research Centre, European Commission, Seville, Spain.

Abstract

This study aims, through a short experimentation, to empirically identify price patterns in popular products from large online retailers. A set of 35 products and prices were monitored for 15 days, three times per day. Three simple price patterns were identified, and four patterns involving two or more sellers were described. The simple price patterns were Temporary rises and fall of prices, Alternation between two prices, and Ladder steps of prices.Compound pattern prices were Price chasing, Price exchange, Mimic at a lower or similar minimum prices, and Conditioned appearance, most of them described in economic literature. This research does not discuss the use of algorithmic pricing when setting prices by online retailer but it could be involved. Next steps in this research consider to wider the number of analyzed products and to increase the frequency and time of their monitoring.

Keywords: *price patterns; algorithmic pricing; pricing technology; online retailers.*

1. Introduction

Pricing technology has recently attracted the attention of academics, practitioners and regulators. Although the use of pricing algorithms has a long history (Calvano et al., 2020), concerns about algorithmic collusion have only recently emerged due to the increased sophistication of learning algorithms and their propensity to uncover collusive pricing rules.

The literature regarding collusive associations and pricing technology is vast. To cite a few, the degree of synchronization in price changes in online markets was studied by Gorodnichenko & Talavera (2016). Cavallo (2018) studied the online competition based on algorithmic pricing technologies on large retailers using different categories of products. Brown and MacKay (2021) studied the price time series of allergy drugs from five online retailers in the United States, and the reaction to price products among rivals.

This study aims through a short experimentation to empirically identify price patterns in popular products from large online retailers. For that pupose, a set of 35 products and prices were monitored during 15 days. Methodology applied is described in Section 2, and main price patterns identified are described in Section 3. Final conclusions are provided in Section 4.

2. Material and methods

To monitor the progress of prices and detect identifiable patterns, prices from 35 consumer products were observed during 15 days (from 5th Febrary to 20th February 2023) in online market places in Spain, three times per day. Products categories were Electronics (20 products, e.g., Apple Iphone 11 64 GB Black), Home (two products, e.g., Philips Wake-up Light HF3500/01), Kitchen (five products, e.g., Cafetera Bialetti Venus, 6 cups), Toys (four products, e.g., LEGO 71043 Harry Potter Hogwarts Castle Model) and Handicraft (four products, e.g., Fimo Soft Modelling Clay, Lemon, 57 g). Online market places selling these products were obtained from Google Shopping, and later these products located on these markets and the product pages were web-scrapped and parsed the resulting html files. Similarly, the same 35 analyzed products were identified in Amazon online store since was not provided by Google Shopping. Then, products were matched in Amazon according to their similar name, characteristics and price of those analyzed products in Google Shopping. A longitudinal database of 12206 observations, and time, product, online seller and price variables were built after processing each product page. Those merchants renting any of the analyzed products were not considered, as well as those product pages in which the price of the product was temporarily hidden. When an online seller was present multiple times (e.g., Fnac and Mediamarkt, which have multiple online shops in Spain) just two online shops were considered, namely, the one closest to the place from where the research was carried out (Seville, Spain) and the one with the lowest price of the product, most of the times coinciding

in price (e.g. uniform pricing) (Cavallo, 2018). Some sellers were found to be associated with other merchants like eBay.

3. Results and discussion

Analyzed products were sold by 156 online sellers, and as expected, the higher frequency of change of process were found in Electronics category (results not shown). Some recurrent dynamic of price changes were analyzed by sellers when selling the same product over time, and some common pattern were detected. Some of these patterns was identified for a single merchant (simple patterns), and in conjunction with others (composite pattern). Most of the observed price patterns can be present individually or in combination. Next is illustrated and described those types of patterns.

3.1. Simple patterns

These patterns were detected in Amazon, asgoodasnew.es, ebay.es and Acelstore, mostly on mobile phones.

(1) **Temporary (and drastic) rises or falls**. This happens when product prices are suddenly rised or dropped to later return at the prior or similar price within hours. In Figure 1 (left), it is illustrated a 46.6% increase in the price of Amazon's Samsung Galaxy S22 Ultra SM-S908B 17.3 cm. At the center, the price of Sony Playstation 5 Standard Edition - 825GB White was 50.9% raised by ebay.es. In this latter case, prices before and after the rise are different. The magnitude of the rise or drop can be very varied, as in the Figure 1 (right) in the case of Apple iPhone XR Black 64 GB sold by ebay.es, with a 25.2% price drop. This price dispersion in online retailers was described by Duch-Brown & Martens (2014) but in a longer period of time and not on a hourly basis like in this study.

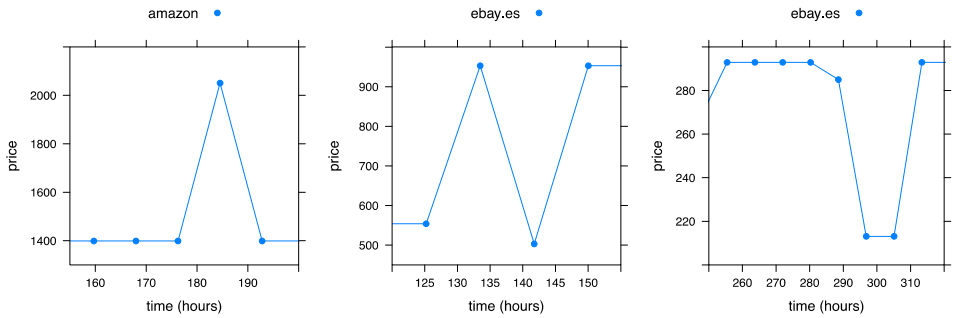


Figure 1. Occasional dizzying rises and falls

(2) **Alternation between two prices**. The price of the Apple iPhone 8 64GB Gray alternates between two set prices, that can vary over time. In Figure 2 (left) the price was set to 260.9€

and 438€ by ebay.es. On the right, and for the same product, asgoodasnew.es set the prices initially at 195€ and 245€, and some hours laters at 175€ and 225€. The range of prices is constant in Figure 2 (right).

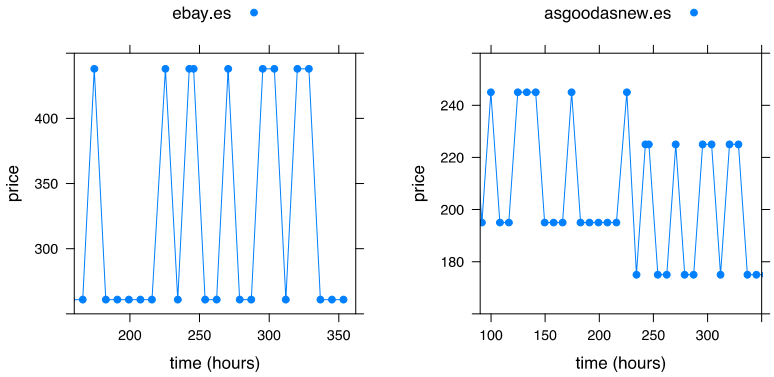


Figure 2. Alternation between two prices

(3) **Ladder steps**. This dynamic reminds the steps of a ladder, with variation in the height and wide (time) of the step. Figure 3 (left) shows this pattern for the Apple iPhone 11 64 GB Black sold by Acelstore. The prices ranged from 317€ to 377€ in three steps, two of them at 327€ and the last one at 333€. At the center, the price of Sony XDRS41DB.EU8 – Portable Digital Radio sold by Amazon begins at 77.99€ and ends at 98.11€, with two intermediate steps at 87€ and 90€.

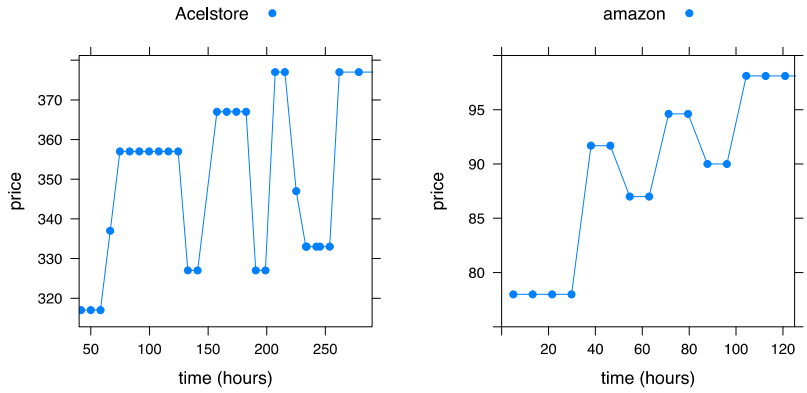


Figure 3. Staircase with different height steps

3.2. Compound patterns

These price dynamics imply at least to two sellers selling the same product, and in some cases can be observed a combination of them, as in the simple patterns.

(1) **Price chasing.** One seller prices like the other seller, or approximately. In Figure 4, Amazon and Mediamarkt price DJI Mini 2 Fly More Combo (Dron) with the same value during 15 days, except at particular occasions, in a raising trend. This trend was no observed in a descensing fashion in neither of the products observed in this study. For the sake of clarity, another MediaMarkt shops was omitted in the graph, with the same dynamic as its homologous.

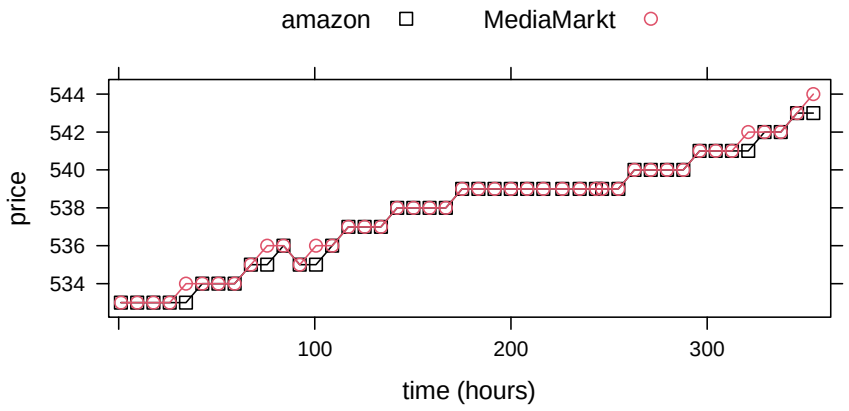


Figure 4. Price chasing

(2) **Price exchange**. Two players exchange the price at a given time. This is illustrated in Figure 5 at prices 959.9€ and 1099€ for the Samsung Galaxy Z Flip4 5g 128GB Grey sold by MediaMarkt and the manufacturer of the mobile telephone.

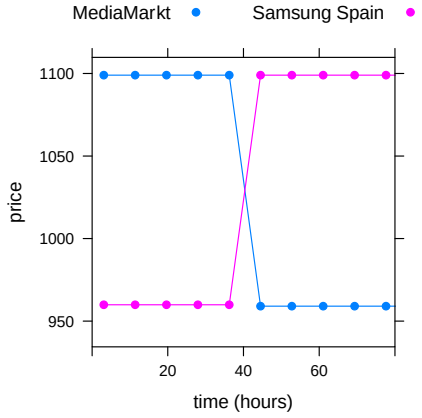


Figure 5. Price exchange

(3) **Mimic at a lower or similar minimum prices**. Two sellers show the same dynamic of prices, but with a different minimum price. In Figure 6 (left), the price for Apple iPhone 8 64GB Gris sold by asgoodasnew.es and ebay.es show different range, but the pattern is similar. One of the seller keep lower prices than the other one. On the right, the difference between the minimum prices for the iPhone 12 128GB Black Apple is lower than in the previous case. This could be referred as a low-price matching practice in literature (Deck & Wilson, 2000). Figure 6 (left) shows a combination of the simple patterns 1 and 3 (Temporary and drastic rises and falls, and ladder steps).

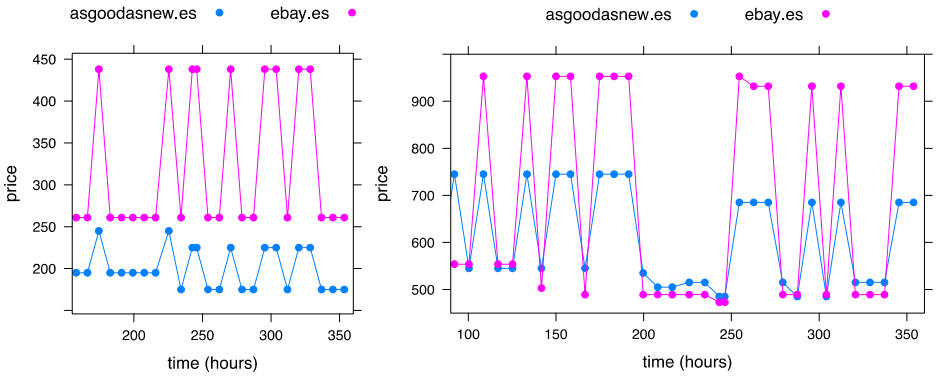


Figure 6. Mimic at different or similar minimun prices

(4) **Conditioned appearance**. The intermittent appearance on some sellers could be conditioned by the presence of other sellers. This could be the case of Fnac, ShopDutyfree.es and Swappie on the Apple iPhone 12 - Black, 128GB. In few occasion the three sellers appear simultaneously.

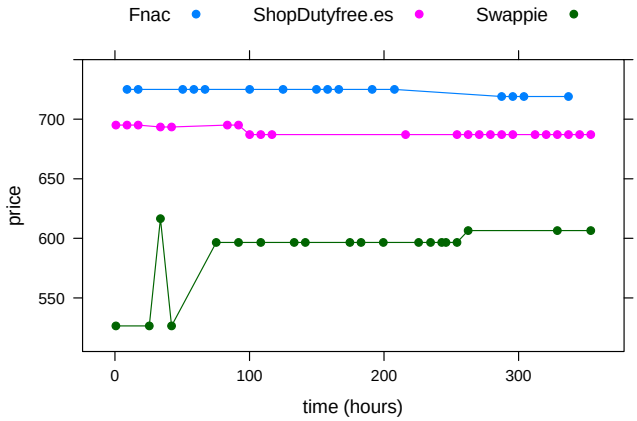


Figure 7. Conditioned appearence

These described patterns are a simple empirical description of the observed price dynamics for some of the selected products for this study. Therefore it is not considered possible causes conditioning such dynamics, as stock status, local seasonal changes, special offers on products, trade rules or collusive agreements among sellers which are difficult to evaluate, to name just a few. Also the human intervention or the automate price-setting using sotware tools (algorithmic pricing) are not assessed. However, patterns described in this work as the compound pattern 1 (price chasing) make difficult to believe an human decision-maker behind that pattern, given the high frequency of price changes within hours. Sample of products selected for this study are popular, but few in number. In the simple described patterns, more than one example has been choosen to illustrate the price dynamic, however, just one in the case of the compound ones, since it should require to wider the span of the research for finding the compound patters in repeated occasions to agreed them. Further research includes to extend the observational period of prices, extend the list of products to be observed over time, and increase the observation frequency of prices.

4. Conclusions

This empirical study aimed to describe observational patterns of prices in online platforms. For that purpose, 35 popular products sourced using Google Shopping were scrapped from online market places and monitored during 15 days. Three simple patterns were detected and two examples of each provided (e.g., temporary rises and fall of prices, alternation between two prices and ladder steps of prices). Besides, four price patterns implying two or more sellers were also described (e.g., price chasing, price exchange, mimic at a lower or similar minimum prices and conditioned appearance). Given the high frequency of price changes observed, it seems algorithmic pricing could be operating on price decisions. Next steps in this research consider to wider the number of analyzed products and to increase the frequency and time of their monitoring.

References

Brown, Z.Y., & MacKay, A. (2021). Competition in pricing algorithms. American Economic Journal: Microeconomics (forthcoming). Working paper. SSRN. Available at: https://ssrn.com/abstract=3485024.

Cavallo, A. (2018). More Amazon effects: online competition and pricing behaviours. NBER Working Paper No. 25138.

Calvano, E. et al. (2020). Protecting consumers from collusive prices due to AI. *Science*, 370, (6520), 1040-1042.

Deck, C.A., & Wilson, B.J. Interactions of automated pricing algorithms: an experimental investigation, in Proceedings of the 2nd ACM Conference on Electronic Commerce, 77–85.

Duch-Brown, N., & Martens. B. (2104). Consumer benefits from the EU Digital Single Market: Evidence from household appliances markets. JRC Technical Reports. Digital Economy Working Papers 2014/03. JRCB9991.

Gorodnichenko, Y., & Talavera (2017). Price setting in online markets: basic facts, international comparisons, and cross-border integration, *American Economic Review*, 107, 249–282.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16419

Estimating policy uncertainty within monetary policy debates

Sami Diaf[1], Florian Schütze[2]
[1] Department of Socioeconomics, University of Hamburg, Germany, [2] Department of Socioeconomics, University of Hamburg, Germany.

Abstract

Studying policy uncertainty contained in collections of documents has been a major task for political researchers and economists, who aim at measuring this degree exclusively with wordlists and topic models to feed further econometric inferences or test hypotheses. Such bag-of-word applications constrain the analysis and cannot render a clear picture of uncertainty drivers and their persistence, even if semi-supervised strategies may offer coherent improvements at the topic level. This work proposes a semantic search strategy, using Top2vec, to identify sources of uncertainty, at the debate level, and uncover coherent topics whose representations will be used to get uncertainty prevalence within each debate. Unlike aggregate-level measurements, this strategy is suited to study per speaker contributions at central banks, where uncertainty is regarded as a forward guidance tool and a key strategy when devising monetary policy actions. Applied to FOMC transcripts (1994-2016), the resulting semantic space yields non-overlapping topic vectors indicating a dominance of economic discussions in uncertainty formation within committee meetings, while risks concerns are bounded to financial markets and investments using an investor jargon. Moreover, results demonstrate the importance of experts' contributions in steering the economic debate, hence coloring uncertainty with words not found in traditional uncertainty wordlists and diffusing a significant persistence to uncertainty prevalence during debates that exhibits fractal patterns.

Keywords: *Uncertainty; Semantic search; Topic models; Monetary policy*

1. Introduction

Narrative economics (Shiller, 2017) popularized the already existing interest in extracting information from data using a variety of techniques, often borrowed from machine learning, and use the results as covariates to augment further inferences (Gentzkow, 2019). Social scientists, as opposed to computer scientists, pay attention to results but not the techniques used, resulting in a dominance of bag-of-word methods when it comes to analyzing text data. This strategy, even if it succeeds in uncovering latent patterns, remains suboptimal and constrained (Grimmer and Stewart, 2013). It overlooks the semantic features stemming from documents (Ash et al., 2021), crucial to understanding narrative signals authors try to send throughout their texts.

Advances in natural language processing investigated the importance of context words and the possibility to transform words into vectors encompassing the semantic and syntactic meaning, also known as distributional representation (Mikolov et al., 2013). This concept was later extended at the paragraph or document level (Dieng et al., 2019) and adopted by other architectures (Angelov, 2020) that make use of a dual word-document embedding to efficiently search for meaningful and coherent representations.

Central banking, as an active economic field of interest, witnessed several contributions belonging to the text-as-data fashion. Mostly to gauge sentiments communicated by available corpora or to scale central bankers (Baerg and Lowe, 2020) for an understanding of potential partisanship among monetary policy members or investigate transparency within these committees (Hansen et al., 2018). However, communication-specific characteristics as for ambiguity (Baerg, 2020) and consensus (Meade and Stasavage, 2008) make it difficult to discern intrinsic features as for uncertainty, which is considered as a forward guidance tool in modern central banking (Greenspan, 2004). The existing indices employed to gauge uncertainty based on word counts (Baker et al., 2016) were found to be informative and able to be augmented with further machine learning techniques to improve their accuracy (Tobback et al., 2018). But they fail to determine sources of uncertainty and tie them to specific topics of interest that better explain the context used for such assertions. Moreover, the availability of sentiment dictionary or wordlists for uncertainty (Loughran and McDonald, 2011) cannot guarantee an effective application as context words, being corpus-specific, are often ignored when building such indices, despite being able to confirm hypotheses and meet economic developments.

Particularly, Federal Open Market Committee (FOMC) transcripts have been widely investigated to assess the communicative content as for transparency (Hansen et al., 2018), scaling members' preferences (Baerg and Lowe, 2020) or assessing objectives (Shapiro and Wilson, 2019). Their public release, although with a five-year delay, came as a transparency effort toward a more public-oriented monetary policy, so to end the long-standing secrecy

that prevailed until the end of the 1980s, in what was qualified as *monetary mystique (*Goodfriend, 1986).

We propose in this paper a semantic, topic-based approach to assess policy-based uncertainty using *Top2vec* algorithm, to identify relevant topic structures and terms semantically related to uncertainty from the collection of FOMC transcripts. While usual probabilistic bag-of-word methods use word frequencies to learn global topic structures, they do not fit corpora with a debate structure, which requires specifications to learn local topics. Moreover, the use of prior information as for pre-trained embedding models might be useful to get coherent topics but comes embedded with an information bias (Papakyriakopoulos et al., 2020) when applied to domain-specific corpora. We argue that semantic search strategies are better suited to uncover semantic uncertainty as explained by (Szarvas et al., 2012) without further requiring post-hoc inferences or prior information from external sources. Nonetheless, it is possible to detect uncertainty origins and study their persistence within each debate to quantify its memory occurrence using *Rescaled Range* (R/S) Analysis (Mandelbrot and van Ness, 1968).

Applied to the corpus of per-speaker FOMC transcripts in the United States (1994-2016), the methodology uncovered a macroeconomic color of uncertainty, stemming from economic debates and forecast-based discussions at the meetings, while risk concerns seem to target mostly financial markets. Policy discussions confirm the consensus feature of central bankers and remain exclusively steered by FOMC members, who are prone to use a more neutral, ambiguous tone than in economic debates.

Our paper makes two distinct, significant contributions. In topic models, we demonstrate the efficiency of semantic search models and their ability to uncover local topics when analyzing debate-structured corpora, freeing the analysis from potential biases arising when overlooking the debate dimension. In policy analysis, we reinforce the *policy context* by searching topics, documents and words related to a specific domain or task from the source, instead of doing it at the word level. This enables us to get an extended representation of semantic uncertainty (Szarvas et al., 2012). Moreover, we introduce the concept of *subject persistence* within debates by studying the statistical properties of persistence measurements as for Hurst exponent (Mandelbrot and van Ness, 1968), to examine the regular use of uncertainty within monetary policy debates.

2. Methodology

Traditional topic models are based on the Dirichlet distribution, which is not sequential and unable to capture time-based topics and unbiased topic prevalence over time. For the case of debates, the use of a dynamic topic model is problematic, as defining a time frame to learn topics could be biased. This is due to debates might be fierce targeting few topics or with a

broad spectrum, spanning over many topics. The relatively small size of debate contributions makes it very difficult to capture the rhetoric used by the speakers when adopting a bag-of-word strategy. We argue that semantic representations make it possible to quantify/study a given subject, whether be it a topic of a keyword, and track its prevalence throughout the debate by computing a persistence index, as for Hurst exponent (Mandelbrot and van Ness, 1968), that informs us about its relative incidence throughout the debate. A fierce debate dealing with a given subject is likely to demonstrate persistence, denoting a permanent prevalence that needs to be captured via a measure, while a vague debate has mostly anti-persistent, short memory patterns. The Hurst exponent, resulting from the Rescaled Range Analysis (Hurst, 1951), is used to measure the degree of variability associated with a given time series. It is linked to a geometric measure of irregular shapes known as the *fractal dimension* (Mandelbrot and van Ness,1968).

For the case of monetary policy practices, debates are not known to be fierce, even if dissents are likely to happen but not explicited (Meade and Stasavage, 2008) either in the resulting transcripts of when casting votes. This reinforces the hypothesis that a subject occurrence within a single debate is likely to have a limited variability that could be gauged via a unique fractal dimension. It results, in this paper, the application of the R/S Analysis within each debate on the computed uncertainty scores for each contribution so to give an unbiased measure of how uncertainty was persistent during the FOMC meetings.

3. Data and Results

FOMC transcripts from 1994 to 2016 were gathered from the Federal Reserve website, consisting of 250 documents featuring mostly meetings transcripts as well as transcribed conference calls. Texts were cleaned and decomposed into individual contributions (54,173 entries), so to stress out the debate dimension and an equal weight for each participant at the meetings. Such a scheme maximizes topic detection and will not overlook small contributions that may appear as distinct topics related to the studied task.

Applying Top2vec to the corpus yields 463 local topics related to monetary policy committee inner functioning, where uncertainty-content is correlated to situations where information is vague, ambiguous, or misleading.

Table 1 shows the 10 most correlated topics with the word *uncertainty*. Topics 1, 3, 8 and 9 comprise technical terms used at presentations, usually conducted by economists related to the economic status, while other topics seem to be related to the international environment (topic 2), housing sector (topics 4 and 7), and economic analysis (topics 5 and 9).

Table 1. Top 10 topics correlated with "uncertainty" and their 10 most related words

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
error	war	tendencies	weak	moderation
bands	terrorist	projections	ui	premia
intervals	invasion	italics	housing	empire
errors	iraq	narratives	household	reassuring
fan	kuwait	assessments	deleveraging	downside
frb	knightian	tendency	labor	trajectory
interval	geopolitical	panels	availability	risk
stochastic	military	your	drags	outlook
width	gulf	column	households	portend
model	scandals	clustered	consumer	benign

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
contacts	burn	fan	okun	unhinged
sixth	resets	charts	law	smidgen
atlanta	tapped	bands	nairu	mishkin
directors	subprime	nondissenting	gap	blips
anecdotal	mortgages	obligated	relationship	bad
retailer	conduits	dissenters	model	knightian
reports	delinquencies	errors	intercept	nimble
optimism	jumbo	histogram	statistical	loop
regards	sheets	width	regularity	pray
reported	uninsured	nonvoters	arguable	very

Source: Authors' own calculations.

The top 10 topics in table 1 are weakly correlated with the word "uncertainty" and the correlation coefficient ranging from 17% to 25%. Table 2 confirms the dominance of the

forecasting jargon, where the top 20 words correlated with the term *uncertainty* are given, along the cosine similarity of their respective vectors, taken from the *Top2vec* semantic space.

Table 2. Top 20 words correlated with the word "uncertainty" and their correlations given by the cosine similarity.

Word	Correlation	Word	Correlation
uncertainties	0.694	regarding	0.550
uncertain	0.659	risks	0.547
surrounding	0.637	face	0.546
confidence	0.624	potential	0.545
many	0.593	there	0.545
around	0.576	outlook	0.545
see	0.565	sense	0.544
about	0.561	of	0.543
given	0.557	still	0.541
considerable	0.554	more	0.540

Source: Authors' own calculations.

The cluster of documents formed by the reduced semantic space will be used to get the correlation of each individual level contribution with the word "uncertainty". In other terms, each speaker in each meeting gets a score that translates the degree of uncertainty used in his/her contribution during the FOMC meetings.

It results in the estimation of uncertainty persistence at each meeting using rescaled range analysis, namely, the Hurst exponent, which provides a good measurement of persistency for time series data. For values of Hurst exponent ranging from 0 to 0.5, the series is said to be an anti-persistent, mean-convergent process, while values from 0.5 to 1 indicate a persistent process that digresses from the mean. A 0.5 value indicates a memoryless process, known as the Brownian motion.

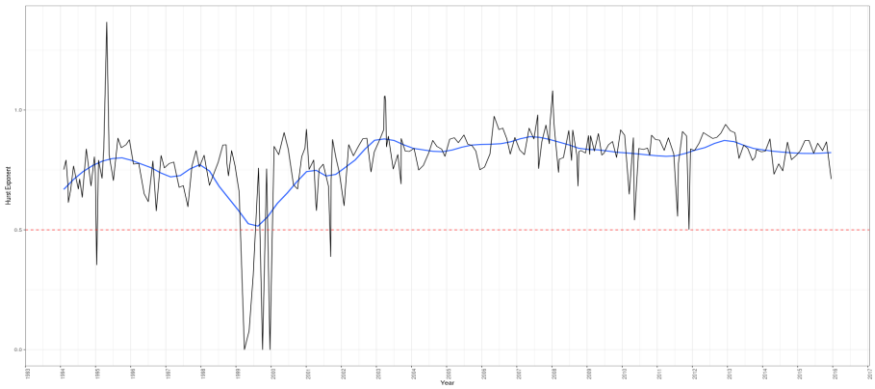


Figure 1. Estimated uncertainty persistence using Hurst exponent within each debate. Blue line is the Loess smoothed curve. Source: Authors' own calculations.

Figure 1 shows an overall erratic, highly volatile persistence, with an anti-persistent epsidode during the period 1999-2001 that could indicate a frequent use of unsure wording/rhetoric in the debates. Overall, Hurst exponent demonstrates a significant occurrence of uncertainty-related contributions, which could be interpreted as a specific consensus shared by some members, or a *herding*, at the contrast of a general *consensus* (Baerg, 2020; Meade and Stasavage, 2008) that was found prevailing at the FOMC meetings. In other terms, a member hinting uncertainty is likely to be followed by others who do the same, therefore leading to a Hurst exponent exceeding 0.5. Rarely, the Hurst exponent records values at 0.3 or below, which could be interpreted as a polarization of the debate around uncertainty.

4. Conclusion

Uncertainty, as an adopted strategy by central bankers, remains difficult to pin down with automated text analysis. Given the necessity to perform such granular tasks, learning local structures via topic vectors helped to identify sources of uncertainty at FOMC meetings, consisting mainly of the use of forecasters' jargon by economists, rather than policymakers. The idea of topic persistence was introduced to gauge the variability of uncertainty-related contributions within each meeting. The Hurst exponent, as a persistency measurement, indicated a clear persistent trend, if we except periods of recession/crises that showed an anti-persistent behavior. These findings confirm consensus being a key feature of modern central banking communication. Latent differences among speakers regarding uncertainty were found to be linked to the economic outlook, usually debated at the opening of the meetings, with the use of proper words related to economic forecasting.

References

Airoldi, E. M., Blei, D., Erosheva, E. A., & Fienberg, S. E. (Eds.). (2014). Handbook of Mixed Membership Models and Their Applications. Hoboken: Taylor and Francis.

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. ArXiv. Retrieved from https://arxiv.org/pdf/2008.09470

Ash, E., Gauthier, G., & Widmer, P. (2021). RELATIO: Text Semantics Capture Political and Economic Narratives. ArXiv. Retrieved from https://arxiv.org/pdf/2108.01720

Baerg, N. (Ed.). (2020). Crafting consensus: Why central bankers change their speech and how speech changes the economy. New York, NY, United States of America: Oxford University Press.

Baerg, N., & Lowe, W. (2020). A textual Taylor rule: estimating central bank preferences combining topic and scaling methods. Political Science Research and Methods, 8(1), 106–122. doi:10.1017/psrm.2018.31

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. The Quarterly Journal of Economics, 131(4), 1593–1636. doi:10.1093/qje/qjw024

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. J. Mach. Learn. Res., 3(null), 993–1022.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. doi:10.18653/v1/N19-1423

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2019). Topic Modeling in Embedding Spaces. Retrieved from https://arxiv.org/pdf/1907.04907

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. Journal of Economic Literature, 57(3), 535–574. doi:10.1257/jel.20181020

Goodfriend, M. (1986). Monetary mystique: Secrecy and central banking. Journal of Monetary Economics, 17(1), 63–92. doi:10.1016/0304-3932(86)90006-1

Greenspan, A. (2004). Risk and Uncertainty in Monetary Policy. American Economic Review, 94(2), 33–40. doi:10.1257/0002828041301551

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267–297. doi:10.1093/pan/mps028

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*. The Quarterly Journal of Economics, 133(2), 801–870. doi:10.1093/qje/qjx045

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. Transactions of the American Society of Civil Engineers, 116(1), 770–799.

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65. doi:10.1111/j.1540-6261.2010.01625.x

Mandelbrot, B. B., & van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. SIAM Review, 10, 422–437.

Meade, E. E., & Stasavage, D. (2008). Publicity of Debate and the Incentive to Dissent: Evidence from the US Federal Reserve. The Economic Journal, 118(528), 695–717. doi:10.1111/j.1468-0297.2008.02138.x

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv. Retrieved from https://arxiv.org/pdf/1301.3781

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in word embeddings. In M. Hildebrandt (Ed.), Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 446–457). doi:10.1145/3351095.3372843

Shiller, R. J. (2017). Narrative Economics. The American Economic Review, 107(4), 967–1004. Retrieved from http://www.jstor.org/stable/44251584

Szarvas, G., Vincze, V., Farkas, R., Móra, G., & Gurevych, I. (2012). Cross-Genre and Cross-Domain Detection of Semantic Uncertainty. Computational Linguistics, 38(2), 335–367. doi:10.1162/COLI_ a_ 00098

Tobback, E., Naudts, H., Daelemans, W., Junqué de Fortuny, E., & Martens, D. (2018). Belgian economic policy uncertainty index: Improvement through text mining. International Journal of Forecasting, 34(2), 355–365. doi:10.1016/j.ijforecast.2016.08.006

Density forecasts with quantile autoregression with an application to option pricing

Johannes Bleher[1], Thomas Dimpfl[1,2], Sophia Koch[1,2]
[1]Computational Science Hub, University of Hohenheim, Germany, [2]Department of Business Mathematics and Data Science, University of Hohenheim, Germany.

Abstract

This paper presents a method for estimating the conditional and joint probability densities of multiple random variables using quantile regression, established by Koenker and Bassett (1978), for which the statistical inference has been extended to the field of time series analysis by Koenker and Xiao (2006). We provide a simple and robust framework for estimating auto-regressive, conditional densities, allowing for inference not only on the conditional density itself but also on functions of the modeled random variables, such as option prices. In our application, we demonstrate theoretically, via a simulation study and in out-of-the-sample density forecasts the effectiveness of our approach in estimating option prices with confidence bounds implied by the estimation method. Our findings suggest that quantile autoregression is effective in forecasting conditional densities and can be used for option pricing. The flexibility of our method in incorporating conditioning information, such as past returns or volatility, has the potential to further improve forecasting accuracy.

Keywords: *Quantile Regression, Conditional Density Forecasts, Option Pricing*

References

Koenker, R., & Bassett, G., Jr. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.

Koenker, R., & Xiao, Z. (2006). Quantile autoregression. *Journal of the American Statistical Association, 101*(475), 980–990.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16482

Spatial distribution of health care facilities in City of Cape Town, South Africa

Sebnem Er[1]
[1]Statistical Sciences Department, University of Cape Town, South Africa.

Abstract

Health care in South Africa is the fourth largest item of government expenditure. Most South Africans do not have medical insurance and therefore seek medical care from public health care facilities such as public hospitals, clinics and community day centres. In City of Cape Town (CoCT), in 2016 there were 149 health care facilities and in 2023 this number increased to 160 facilities with an annual growth rate of 1.19%. The population in 2016 was 4 million and in 2021 it is estimated as approximately 4.76 million which indicates an annual growth rate of 3.5% between 2016 and 2021. The annual growth rate of health care facilities (primarily representing data applicable to the City of Cape Town's Environmental Health Department) is clearly not matching the annual growth rate of the population which is potentially a big problem in a fast growing region like CoCT. The main aim of this paper is to analyse the distribution of health care facilities (clinics, hospitals) within the City of Cape Town, South Africa using spatial kernel density estimation methods and explore what factors affect the number of health care facilities within each ward using generalized linear models.

Keywords: *Health care facilities; spatial distribution; geographically weighted poisson regression.*

1. Introduction

City of Cape Town (CoCT) is the only metropolitan municipality in the Western Cape province of South Africa that is the home for the Mother City, Cape Town. The municipality is made up of 111 wards and as of the latest available 2011 census data, the total population of the municipality is 3,740,026. According to city's measures the population in 2016 was 4,005,016 and in 2021 it is estimated as approximately 4,758,433. The very recent census results are not published yet however it is in no doubt that the city is growing fast with an annual growth rate of 1.38% from 2011 to 2016 and 3.5% from 2016 to 2021. One of the main challenges that South Africa faces is the access to health care services, most people do not have a medical insurance and they seek the services from public health care facilities. Health care facility point pattern distribution in 2021 is provided in Figure 1, where the point pattern plotted over the population measures within wards is provided on the left and on the right pane, the pattern is plotted over the total number of health care facilities within each ward. It is clear that highly dense areas have fewer number of health care facilities compared to less populated areas. Considering that most South Africans take public transport to get to work and to access the services provided within the municipality, it is crucial to explore the distribution of health care facilities within the CoCT.

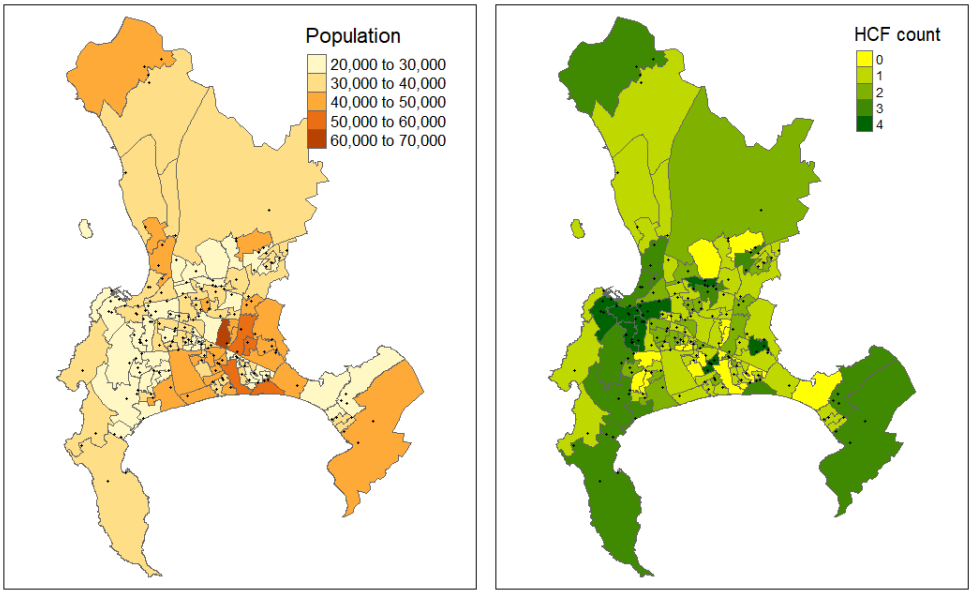


Figure 1. CoCT health care facility distribution overlaid on Left: CoCT population distribution in 2011 (census data), Right: CoCT ward level health care facility count.

In this paper, the intensity of the health care facilities will be estimated using kernel smoothing estimation method and thereafter the total number of health care facilities within 111 wards will be modeled using generalized linear models. There are in total 160 health care facilities, of which 70 of them are clinics, and 48 of them are hospitals (district, regional, psychiatric, tertiary and private hospitals). The data is obtained from City of Cape Town's open data portal which provides publicly accessible data. "*Access to City information helps to increase transparency, as well as benefit the wider community and other stakeholders*" (CoCT Open Data Portal). There are very few studies making use of the data available and analysing with appropriate methods. This research aims to use the health facility data to provide insight on the disparity of the distribution of the facilities within the city.

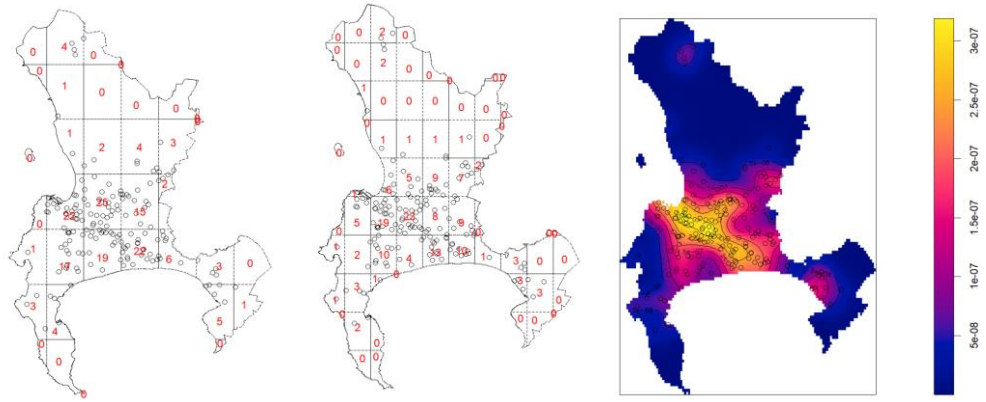
2. Spatial Kernel Density Estimation

Spatial point patterns can be completely random, clustered and regular within a bounding region (A). A completely spatial random (CSR) point process, often associated as homogeneous poisson process (HPP), asserts that the number of events in a quadrat a_j with area $|a_j|$ follows a Poisson distribution with mean $\lambda|a_j| = \frac{n(X)}{|A|}|a_j|$ where λ is the expected number of points within region A and $\bar{\lambda} = \frac{n(X)}{|A|}$ is the unbiased estimate of the true intensity. The second assumption asserts that given n events x_i in a region a_j, the x_i are an independent random sample from the uniform distribution on a_j (Cressie, 1991, pp.586).

There are many different measures that can be calculated to test for completely spatial random processes. Quadrat counting and the associated chi-square test, distance methods and second order statistics are a few of many ways to explore divergence from CSR. In most cases, such as the distribution of the health care facilities, it would be naïve to assume CSR and therefore best practice would be to estimate a varying intensity rather than a constant one. The estimation of the intensity that varies over the location ($\hat{\lambda}(x)$) is known as the inhomogeneous poisson process (IPP) which is a generalisation of CSR. For IPP, the estimation of the intensity can be done by means of non-parametric kernel smoothing or by means of a parametric function for the intensity whose parameters are estimated by maximising the likelihood of the point process. Non-parametric kernel smoothing estimator is provided in the following equation (Bivand et al., 2008, pp.165):

$$\hat{\lambda}(x) = \frac{1}{h^2}\sum_{i=1}^{n} \kappa\left(\frac{\|x - x_i\|}{h}\right)/q(\|x\|)$$

where $\kappa(u)$ is a bivariate and symmetrical kernel function (such as quartic - spherical kernel) and $x_1, x_2, \ldots, x_n$ are the data points. $q(\|x\|)$ is the border correction to compensate for the missing observations that occur when x is close to the border of the region A. In this paper, border correction is not a concern since the entire region (CoCT metropolitan municipality) is under study. Finally, the bandwidth h measures the level of smoothing where small values will produce very peaky estimates, and large values will produce very smooth functions. CoCT health care facility intensity is explored in **R** (R Core Team, 2022) with **spatstat** package (Baddeley, et al., 2015) using 7x7 and 10x10 quadrats and corresponding chi-square tests. The results from quadrat tests are provided in Figure 2 left and middle panes. According to the results, it is evident that the intensity is not constant across the study region and therefore it is important to estimate the intensity that varies over location. The estimation of health care facility intensity varying over the study region has been done using kernel density estimation with a bandwidth optimized using cross validation of the minimization of mean-square error criterion formulated in Diggle (1985).



$\chi^2 = 222.12$
$(p - value \ll 0.001)$

$\chi^2 = 294.76$
$(p - value \ll 0.001)$

Figure 2. CoCT health care facility quadrat counts and chi-square test results for Left: 7x7 quadrats, Middle: 10x10 quadrats. Right: CoCT health care facility kernel density estimation.

3. Geographically Weighted Poisson Regression

To model the health care facility counts within wards, Poisson regression specific to count data will be utilised. Since the number of health care facilities are different locally, geographically weighted regression models are investigated. Geographically weighted poisson regression (GWPR) is the locally predicted version of a Poisson regression model which involves the selection of a bandwidth chosen by LOOCV or manually for count data. Once a decision is made on the bandwidth, the model is fit with a local kernel function

which is usually a Gaussian kernel function. The results simply are the estimated geographically varying coefficients within the chosen bandwidth provided as follows (Fotheringham et al., 2003):

$$y_i = e^{\beta_0(u_i,v_i)+\sum_k \beta_k(u_i,v_i)x_{k,i}}$$

where $\beta_0(u_i, v_i)$ is the intercept parameter specific to location i and $\beta_k(u_i, v_i)$ are the coefficients of independent variables at location i. Estimation procedure is conducted in **R** (R Core Team, 2022) using **GWmodel** package (Gollini et al., 2015, Lu et al., 2014).

Table 1 provides the single variable GWPR results for predicting health care facility counts in a ward using several variables such as (i) white population ratio, (ii) black African population ratio, (iii) population density of each ward, (iv) ratio of employed population to unemployed population, (v) unemployed population ratio, and (vi) employed population ratio.

It has to be noted that all models need to be assessed with extensive caution since the pseudo R-square values obtained between the predicted and observed number of health care facilities are very low. Keeping in mind the very low measures, it can be seen that Models 1 and 2 indicate that black African population has access to fewer health care facilities compared to white population. When population density per ward is considered (Model 3), it is observed that denser areas have fewer number of health care facilities compared to less dense areas outlining a negative relationship. Looking at Models 4, 5 and 6, employment plays a positive role in the number of health care facilities. The single variable results point out a possible imbalance of the access to health care facilities at ward level. Several other factors such as closeness to a major road or the specialisation of the health care facilities should be considered for a full picture. The model with the minimum AIC measure is Model 5 where the ratio of unemployed to the population is considered. The parameter estimate is negative which indicates that the less the unemployed ratio is the more there is health care facilities.

Table 1: Single variable generalized poisson regression model results

	Model0	Model1	Model2	Model3	Model4	Model5	Model6
Intercept	0.3656 [4.625] (0.000)	0.2148 [2.123] (0.034)	0.5852 [5.334] (0.000)	0.6585 [5.852] (0.000)	0.1957 [1.772] (0.0764)	0.9012 [6.145] (0.000)	-0.4645 [-1.339] (0.181)
White_ratio		0.7395 [2.737] (0.006)					
Black_Afri_ratio			-0.6276 [-2.588] (0.009)				
popdensity				-0.00005 [-3.194] (0.00141)			
Employed/unemployed					0.0245 [2.420] (0.0155)		
Unemployedratio						-5.6113 [-3.924] (0.000)	
Employedratio							2.3221 [2.515] (0.012)
bw	25	37	105	38	32	102	97
AIC	108.27	103.22	103.03	98.38	104.91	93.96	104.09

bw: bandwidth (# of nearest neighbours), AIC: Akaike information criterion

Table 2: Single variable GWPR model results with minimum, maximum, median and 1st and 3rd quartile values for the specific variables

	Model0 Intercept	Model1 White_ratio	Model2 B_A_ratio	Model3 popdensity	Model4 Emp/Une	Model5 Unempr	Model6 Emplr
Min	0.0864	0.4767	-0.6290	-0.00006	0.0111	-6.0797	2.1592
1st Quartile	0.1770	0.6939	-0.6159	-0.00005	0.0238	-5.9119	2.5107
Median	0.3061	0.9653	-0.6125	-0.00004	0.0392	-5.7813	2.6779
3rd Quartile	0.3911	1.3554	-0.6084	-0.00003	0.0528	-5.6277	2.7594
Max	0.5162	1.6092	-0.6004	0.00000	0.0616	-5.2767	2.8350
Correlation	0.3910	0.4236	0.2662	0.4362	0.4279	0.4183	0.2732
Pseudo-R square	0.1529	0.1797	0.0709	0.1903	0.1831	0.1750	0.0746
AIC	103.75	99.388	103.11	97.64	99.87	93.87	103.61

4. Conclusion and Future Work

The aim of this research is to provide insight on the distribution of the health care facilities within the City of Cape Town using publicly available data from the city's data portal and to draw attention to the fact that there is a disparity in the distribution of health care services. With this aim in mind, the health care facilities in the city have been initially analysed in an exploratory manner using quadrat tests and it has been concluded that the intensity of the health care facilities is not completely spatial random. The varying intensity is modeled with a kernel density function and it has been observed that the health care

facilities are localized in certain areas in CoCT. In order to model the health care facility counts within each ward and to examine the effects of potential variables, different geographically weighted poisson regression models were estimated. It has been seen that there might be a possible imbalance in the distribution of the facilities within the metropolitan city especially where black African population has access to fewer health care facilities that are nearby compared to white population. Previous studies have also found similar imbalances in the locations of health care facilities (Lee, 2013). Lee (2013) analysed the distribution of health care facilities located in the metropolitan city of Daejeon, South Korea and found that there is a disparity. In City of Cape Town, given that the number of nearby health care facilities are limited to certain areas, the findings of the research could be used by policy makers to improve the distribution of the health care facilities, especially considering that most South Africans who live in the densely populated areas take public transport to get to work and to access the services provided within the municipality. This could be achieved by for example locating more mobile clinics or health care facilities that can provide services for those diseases that are more prevalent within the densely populated areas so that people who need these services will not need to travel far distances to get health care services.

The analysis can further be extended with hot-spot analysis and inclusion of other variables such as the specialty of the health care facility, convenient road access of the health care facility, the distance to city center or a major road, and poverty and socioeconomic status of each ward. More insight from several variables will extend the findings and increase the predictive power.

References

Baddeley A, Rubak E, Turner R (2015). Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, London. https://www.routledge.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/p/book/9781482210200/.

CoCT Open Data Portal. (2023). URL: https://odp-cctegis.opendata.arcgis.com/datasets/cctegis::health-care-facilities-clinics-hospitals/explore?location=-33.867164%2C18.629850%2C10.45

Diggle, P.J. (1985). A kernel method for smoothing point process data. *Applied Statistics Journal of the Royal Statistical Society*, Series C, 34, 138–147.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Gollini I, Lu B, Charlton M, Brunsdon C, Harris P (2015). "GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models." *Journal of Statistical Software*, 63(17), 1–50. doi:10.18637/jss.v063.i17.

Lee, Kwang-Soo (2013). "Disparity in the spatial distribution of clinics within a metropolitan city". *Geospatial Health*, 7(2), 199-207.

Lu B, Harris P, Charlton M, Brunsdon C (2014). "The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models." *Geo-spatial Information Science*, 17(2), 85–101. doi:10.1080/10095020.2014.917453

Municipalities of South Africa, CoCT demographic data, (2023). URL: https://municipalities.co.za/demographic/6/city-of-cape-town-metropolitan-municipality

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Western Cape Provincial, Socio-economic profile for City of Cape Town, (2021). URL: https://www.westerncape.gov.za/provincial-treasury/files/atoms/files/SEP-LG%202021%20-%20City%20of%20Cape%20Town.pdf

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16422

Analysing ride behaviours of shared e-scooter users – a case study of Liverpool

Yuanxuan Yang[1], Susan Grant-Muller[1]
[1]Institute for Transport Studies, University of Leeds, UK

Abstract

The shared e-scooter is a relatively new form of Micromobility service in urban transit. A better understanding of the use of the scheme will help operators and stakeholders promote this travel mode, contributing to a more sustainable, resilient, environmentally friendly and inclusive transportation system. The availability of high resolution sensor-based location data, when co-analysed with socio-demographic survey data allows insights on where, how, and by whom the service is used. This study focuses on analysing the usage pattern of a recently introduced shared e-scooter scheme in Liverpool, UK, combining survey data of users' sociodemographic attributes and their full trip records at a fine spatiotemporal granularity. Recency-Frequency (RF) segmentation is used to categorise user behaviour based on their frequency and recency of usage, and a Functional Signatures (FS) dataset is used to enrich contextual information on the origin and destination of e-scooter trips. Overall, this study provides insights into the behaviour of users of shared e-scooters and how the behaviours might vary in different user groups regarding sociodemographic characteristics. The developed analysis framework is also readily transferable to other cities.

Keywords: *Micromobility; sustainable transportation; e-scooter; location data; customer segmentation.*

1. Introduction

Shared e-scooters are a relatively new form of transport mode in the Micromobility family, they refer to electric scooters that are available for rent through a sharing scheme. Access to the scooters depends on whether the scheme is based on a docked or dockless system. For a dockless system, users can first find and unlock the e-scooter with a smartphone app, ride to the destination, and then leave the vehicle at the destination (sometimes with restrictions or geofencing) for the next user. The benefits of a dockless system (Yang et al., 2019) is that scooter availability is not limited to the fixed points of docking stations, therefore shared e-scooters are flexible to use. They are suitable for relatively short travel and solving the "first/last" mile problem – the distance between public transport station and destination (Yang et al., 2019; Hosseinzadeh et al., 2021).

The shared e-scooter scheme offers a convenient and relatively low-cost travel option, and it can bring various benefits to cities and their inhabitants, including reducing congestion, improving air quality and increasing accessibility to various services (Abduljabbar et al., 2021). Shared e-scooters have recently been introduced in several cities and towns in the UK, under government-approved trials (Speak et al., 2023).

Much research has used a qualitative approach or questionnaire survey (König et al., 2022; Speak et al., 2023) to understand the underlying driven factors or related sociodemographic characteristics to different opinions and usage (self-reported) of e-scooters. Studies that consider actual riding behaviours (as revealed through the high-resolution scooter location data), linked to sociodemographic characteristics are rare. This is despite the potential advantages of the insights generated, such as the choices made by user sub-segments, but is at least partly due to limited access to both data in the literature.

This study analysed the usage patterns of shared e-scooter users in Liverpool, a city in the northwest of England, UK. Users' sociodemographic information from a survey and their full trip records at a fine spatiotemporal granularity are obtained (with their permission) and investigated. This research demonstrates the value of integrating the two data types – a large scale digital database and more traditional user survey, and analyses riding behaviours. It provides evidence on how the behaviours might vary in different user groups (e.g. differences in car ownership).

2. Data and Method

This study focuses on a survey dataset of shared e-scooter users and corresponding journey profiles from a shared e-scooter operator (Voi) in the UK. This research focuses on a subset of records relating to the city of Liverpool, as a case study, though in a future paper we will present findings from a larger number of cities and users.

The survey data includes participants' varying sociodemographic attributes, covering age band, gender, ethnicity, self-reported basic health status, car ownership, household income, employment, occupation, educational attainment, and responses to several questions related to subjective wellbeing. As a part of the survey, participants are asked if they agree to share their trip records for research purposes. With the users' consent, this study obtained 89 participants' (in Liverpool) complete history of shared e-scooter usage (from their first-time use to October 27, 2021). The trip data include the following variables: User ID, trip origin coordinates, trip destination coordinates, trip start time, trip end time, travel distance (route distance) and ride speed.

Recency-Frequency (RF) segmentation is utilised to disaggregate users' behaviours from their complete trip profiles. RF analysis is a technique that has many implications in marketing (McCarty & Hastak, 2007; Beecham & Wood, 2014), and it can be used for segmenting customers based on two factors:

- Recency: how recently a customer has made a purchase or used a product
- Frequency: how often a customer make a purchase or use a product.

Both factors are considered to be good predictors of future engagement (usage or purchase) (McCarty & Hastak, 2007; Beecham & Wood, 2014). To perform RF analysis, customers are ranked for the two factors on a *n*-level scale (e.g. n= 3, 4, 5 …). The scores are further concatenated to give at most n*n segments (Beecham & Wood, 2014).

RF segmentation has rarely (if any, to authors' knowledge) been used for analysing e-scooter data. In the e-scooter dataset, "Recency" scores were determined by the most recent e-scooter trip, and assigning discrete scores with three equal recency bins, from most (score 3) to least (score 1) recent (Beecham & Wood, 2014). "Frequency" scores are calculated in two steps: (1) Calculate the first and last trip in each user's trip profiles, getting the length of the "active" period. (2) the "Frequency" score can be obtained by dividing the total number of trips by the length of the "active" period. People with the highest scores in both Recency and Frequency (3-3) are the most "heavy" and "loyal" users - they ride shared e-scooter frequently, with very recent trips. Those classified as the lowest RF group (1-1) may use e-scooter after registering, but have made relatively few trips afterwards (Beecham & Wood, 2014).

The Functional Signatures (FS) dataset in Liverpool (Samardzhiev et al., 2022) was also applied to enrich contextual information on the origin and destination of e-scooter trips. "*The FS are contiguous areas of a similar urban function with fine spatial granularity. Rich datasets, including census, remote sensing, and point of interest data, were used as inputs for grouping based on a clustering approach* (Samardzhiev et al., 2022)". FS provides several dimensions (or qualifiers) to describe the function of each small area in the UK. In this study, due to a relatively limited sample size and geofenced service area, certain types of FS are

combined. While retaining its qualifiers in terms of service and use (e.g. residential, employment), the density dimension is reduced; for example, "Residential – Low density" is recoded as "Residential". The recoding strategies are shown in Table 1. More details of each type of FS are available in the work of Samardzhiev et al. (2022). FS-related origin-destination (O-D) pairs are discussed in section 3.2. Not all FS types (Samardzhiev et al., 2022) exist in the study area; for example, there is no "Countryside" FS in the shared e-scooter service area in Liverpool.

Each e-scooter trip's origin and destination are intersected with FS boundaries; hence, both origin and destination have corresponding recoded FS type information. With the enriched contextual information, it is possible to deepen the understanding of the trip's purpose. For example, a trip from industrial FS to residential FS may have the purpose of going home.

Table 1. Recoded Functional Signatures types in the case study area.

Family	FS type	Recoded FS type
Industrial	Industrial – Construction site	Industrial
	Industrial - Commercial	Industrial
	Industrial - Manufacturing	Industrial
Residential	Residential – Low density – Well-served	Residential – Well-served
	Residential – Well-served	Residential – Well-served
	Residential - Mixed-use	Residential - Mixed-use
	Residential – Low density	Residential
	Residential	Residential
	Residential Greenspace	Residential Greenspace
Service	Service - Mixed – Low density	Service - Mixed
	Services - Leisure and Cultural	Services - Leisure and Cultural
	Services - Transport and distribution hubs	Services - Transport and distribution hubs
Urban	Urban -Mixed-use – High density	Urban – Mixed use
	Urban - High employment, culture, connectivity	Urban - High employment, culture, connectivity
	Urban - High employment, amenities	Urban - High employment, amenities

3. Findings

3.1. Spatial and temporal pattern

Spatial and temporal patterns of e-scooter trips are explored in this study. The density histogram of trip distance is illustrated in Figure 1, and the most popular travel distance is between 1-1.5 km, showcasing the utility of e-scooters for making relatively short-distance journeys. Shared e-scooters can also be used for longer distance journeys, such as those exceeding 5 km. Figure 2 shows the average trip count (rescaled to [0,1], calculated by the trip starting time) across the 24 hours during the day. 1 indicates the highest hourly trip volume, and 0 indicates no trips were in this hourly interval. During a weekday, there is an evident peak in the afternoon (from 16:00-18:00), also small local peaks around 6:00 and 8:00 in the morning. At the weekend, the curve is more flattened across the day, reaching a relatively high platform from 11:00 to 17:00.

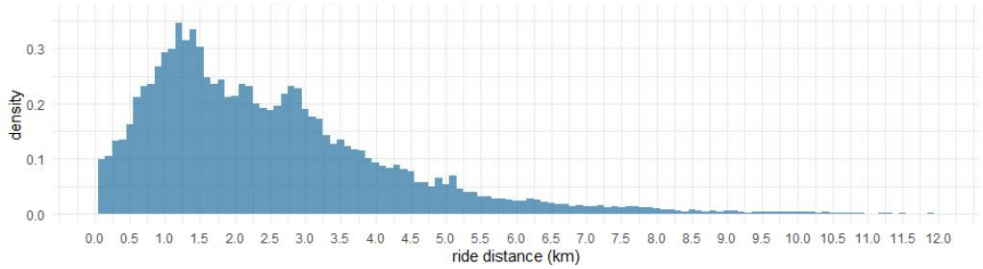


Figure 1. Density plot of ride distance (km).

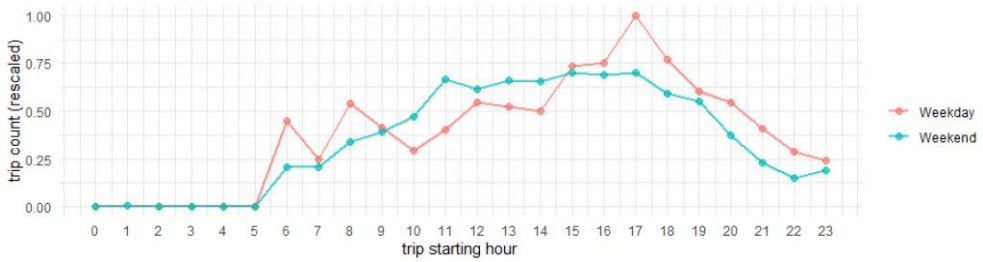


Figure 2. E-scooter trip count by hour (rescaled).

3.2. Result of Recency-Frequency Analysis

RF analysis helps segment e-scooter users into different groups depending on their scores in Recency and Frequency. This study further combined sociodemographic variables with RF scores, and the results are shown in figure 3, using car ownership as an example.

Users in the lowest FS group (1-1, Figure 3, bottom-left) have a relatively low share of "Do not have" a car. With increasing FS scores (2-2, Figure 3, middle-middle), the proportion of non-car owners increased and reached an even share in the highest FS score (3-3, Figure 3,

top right) group. The result implies that e-scooter benefits personal mobility to people who do not have a car, and the service may be considered more helpful and appealing for this group – leading to "loyalty" and favour of using the service. This also aligns with the findings in the literature that shared e-scooter and the wider Micromobility services may be particularly appealing to people who live and travel in urban areas with limited parking options, or owning a car may increase the burden (Bieliński, & Ważna 2020).

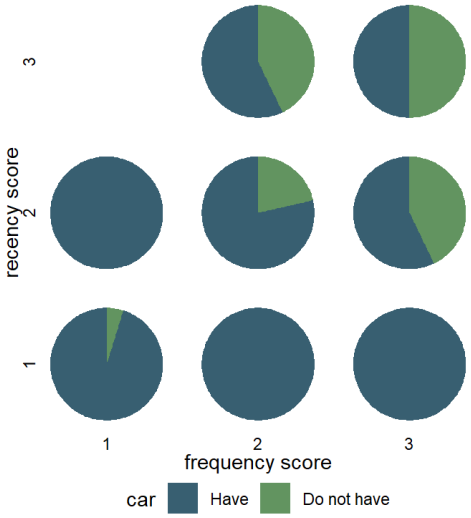


Figure 3. Recency-Frenquency score and users' distribution in car ownership.

3.3. Trip Origin and destination pair

The FS types of e-scooter trip origin and destination are identified, and Figure 4 shows the Sankey plot of O-D pairs of all trips. In total, all "Residential" FS family members have accounted for 49.92% of trip origins and 50.47% of trip destinations (Figure 4). The "Urban – High employment, culture, connectivity" also generated and attracted many travel flows, accounting for 27.50% and 29.16%, respectively (Figure 4). E-scooters are also used for "first/last-mile" trips, linking "Service – Transport and Distribution Hubs" and other FS areas, especially "Urban- High employment, culture, connectivity".

It is also possible to disaggregate the O-D FS types by differentiating sociodemographic attributes such as car ownership. The share of "Residential - Mixed use" as the trip origin is much lower for car owners than the group of not have a car (22.00% compared to 33.83%), and the major difference is contributed by "from Residential – Mixed use to Urban- High employment, culture, connectivity". This suggests that e-scooter provide a favourable alternative mode for non-car owners to travel from "Residential-Mixed" areas to "Urban-High employment, culture, connectivity" FS areas.

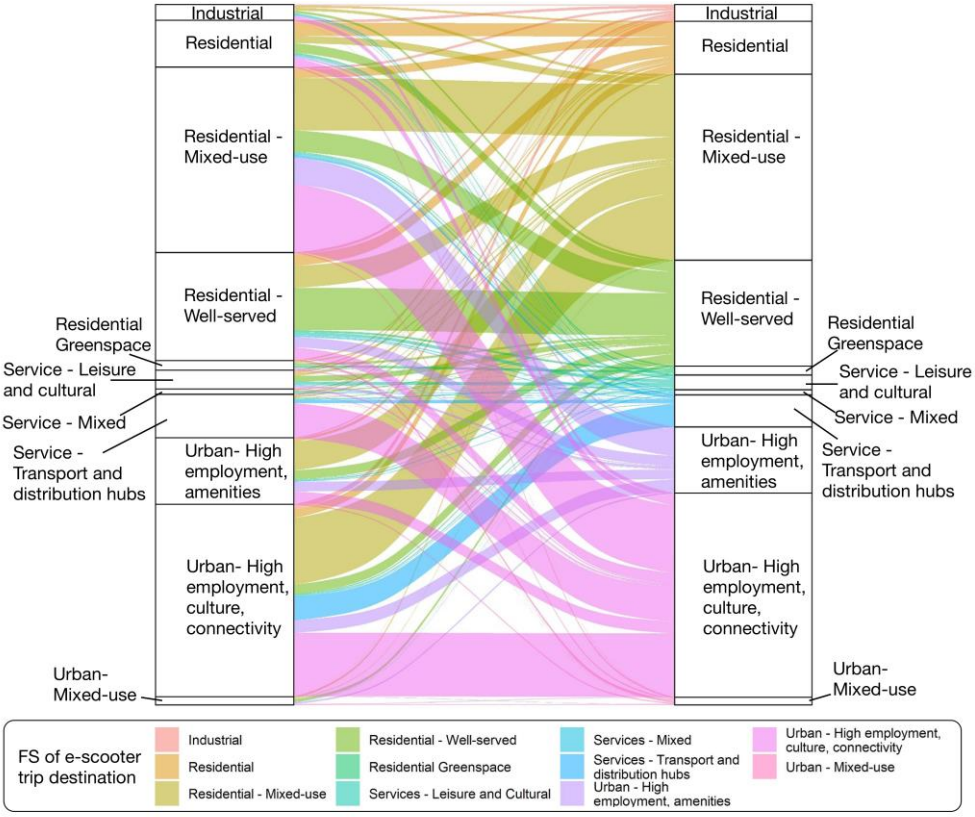


Figure 4. Sankey plot of e-scooter trip origin (left) and destination(right) FS types.

4. Conclusion

This study investigated e-scooter riding behaviours in Liverpool. Different spatiotemporal characteristics and usage patterns are identified and linked to personal sociodemographic characteristics. The findings suggest that e-scooters offer a convenient and flexible short-trip option, and the service is appealing to people who do not have cars. Furthermore, by linking user group segmentation with sociodemographic attributes such as age, gender, ethnicity, educational attainment, and employment, a more comprehensive insight into distinct user groups can be achieved.

This study benefits from rich information covering sociodemographic characteristics and users' full trip records. The trip records are sensor-based and at a fine spatiotemporal granularity, therefore is able to reflect user behaviours in detail, providing evidence on who, where, when, how and why the service is used. Such data at a larger scale also have the potential to reveal the dynamics and rhythm of urban flows in the last-mile. It is worth noting

that this work includes a limited number of samples (participants) in the case study area, and the potential bias issue should not be ignored. The e-scooter travel flow and associated O-D FS types may be impacted by service provision; the availability of scooter and the geofencing of service area could all impact where trips start and end.

The analysis framework utilized in this study can provide a nuanced understanding of user characteristics, encompassing ride behaviors and sociodemographic attributes. This can help identify potential barriers and enable scheme operators and transport management authorities to strategically promote the usage of e-scooters and sustainable travel mode.

Future work might deepen the insights by utilising data at a larger scale (combining observations in other cities), possibly incorporating richer full trajectory data to understand the route choice of different users.

Acknowledgement

This research has been sponsored by the Alan Turing Institute under grant number R-LEE-006. We would like to thank Voi for providing the e-scooter data.

References

Abduljabbar, R. L., Liyanage, S., & Dia, H. (2021). The role of micro-mobility in shaping sustainable cities: A systematic literature review. *Transportation research part D: transport and environment*, *92*, 102734.

Beecham, R., & Wood, J. (2014). Exploring gendered cycling behaviours within a large-scale behavioural dataset. *Transportation Planning and Technology*, *37*(1), 83-97.

Bieliński, T., & Ważna, A. (2020). Electric scooter sharing and bike sharing user behaviour and characteristics. *Sustainability*, *12*(22), 9640.

Hosseinzadeh, A., Algomaiah, M., Kluger, R., & Li, Z. (2021). Spatial analysis of shared e-scooter trips. *Journal of transport geography*, *92*, 103016.

König, A., Gebhardt, L., Stark, K., & Schuppan, J. (2022). A multi-perspective assessment of the introduction of e-scooter sharing in germany. *Sustainability*, *14*(5), 2639.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of business research*, *60*(6), 656-662.

Samardzhiev, K., Fleischmann, M., Arribas-Bel, D., Calafiore, A., & Rowe, F. (2022). Functional signatures in Great Britain: A dataset. *Data in brief*, *43*, 108335.

Speak, A., Taratula-Lyons, M., Clayton, W., & Shergold, I. (2023). Scooter Stories: User and Non-User Experiences of a Shared E-Scooter Trial. *Active Travel Studies*, 3(1).

Yang, Y., Heppenstall, A., Turner, A., & Comber, A. (2019). A spatiotemporal and graph-based analysis of dockless bike sharing patterns to understand urban flows over the last mile. *Computers, Environment and Urban Systems*, *77*, 101361.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16415

Estimation by kernel weighting of parameters related to employment in the confinement period

Beatriz Cobo[1], Luis Castro[1], Jorge Rueda[2]
[1]Department of Quantitative Methods for Economics and Business , University of Granada, Spain, [2]Department of Statistics and Operations Research, University of Granada, Spain.

Abstract

During the period of COVID-19's confinement, new working methods that were not normally used began to be relevant. This is the case of teleworking or the use of new techniques for conducting surveys. The gold standard for carrying out surveys is probability sampling based on face-to-face interviews, but due to this situation of social isolation, non-probabilistic methods, such as online or web surveys, began to be used. However, in order to make reliable estimates from non-probability samples we must use special techniques to reduce the bias that appears in them.

In this paper we will study a technique for bias reduction in non-probabilistic surveys that stands out for its promising results, known as Kernel Weighting. It requires a probabilistic sample as auxiliary information, and its performance can be improved using Machine Learning techniques, such as regularised logistic regression. We will use a non-probabilistic survey focused on studying the employment situation of the Spanish population during COVID-19, and as probabilistic survey the CIS Barometer of May 2020. We will compare the new estimates with those obtained in the original survey, observing important differences.

Keywords: *Regularized logistic regression; kernel weighting; employment; confinement period; COVID-19.*

1. Introduction

The COVID-19 brought with it a situation of confinement never experienced before that affected the habits and behaviors of Spaniards both economically, at work and socially. Focusing on the first Pérez et al. (2022) conducted a non-probabilistic survey using snowball or chain sampling to recruit respondents during the confinement period in Spain and obtained a data set with sociodemographic variables, variables related to residence during confinement with employment status with household chores, with health, and with politics, that it to say, how they lived and felt as well as their perceptions while in lockdown. Cowan (2020) used IPUMS-CPS survey data to examine how workers transitioned between labor market states and which workers have been most affected by the pandemic. Coibion et al. (2020) used the Nielsen Homescan survey to show the large effects of the pandemic on job losses. Cajner et al. (2020) measured the evolution of the US labor market during the first four months of the pandemic using weekly administrative payroll data from the largest US payroll processing company. Fairlie et al. (2020) used CPS data to focus on how the job market crisis due to the pandemic has affected racial and ethnic minorities compared to whites. Schieman et al. (2021) found that the conflict between work and life decreased among people without children at home; in the case of having children, these changes were limited to the age of the youngest child in the home and the degree of work-home integration.

In most of these studies, to carry out an investigation in which we want to know what happens in the population, it is essential to carry out a survey. The ideal in these cases is to use a probabilistic sampling to ensure that the sample is random and we will be able to make inferences later. However, years ago researchers began using non-probability surveys. Non-probability sampling is a method of selecting units from a population using a subjective (ie, non-random) method. Since non-probability sampling does not require a complete survey framework, it is a quick, easy, and inexpensive way to collect data.

Following the onset of the COVID-19 pandemic, conventional survey data collection efforts (e.g., pencil-and-paper surveys as part of population-representative household surveys) came to a halt due to lockdowns, mobility limitations and social distancing requirements. Because of this, what has occurred is an increase in telephone surveys among others to meet the rapidly evolving needs for timely and policy-relevant microdata to understand socioeconomic impacts and responses to the pandemic. Gourlay et al. (2021) provided an overview of options for the design and implementation of telephone surveys to collect representative data from households and individuals. In addition, they identified the requirements for telephone surveys to be used in national statistical offices. De Boni (2020) briefly presented some of the advantages that may have driven web surveys. She talks about the growing popularity of these in the COVID-19 context, since in addition to the possibility of collecting data remotely, it allows social distancing. She also tells us about

the side effects and ethical issues that need to be considered when planning such surveys and interpreting their results. Hlatshwako et al. (2021) tell us about online health research surveys, the challenges in the implementation and interpretation of data from this type of survey, and the considerations that must be made to make the most of the research. Ali et al. (2020) show us how the COVID-19 pandemic has forced researchers to explore innovative ways to collect public health data in an efficient and timely manner. Social media platforms were explored as a recruiting tool for researchers in other settings; however, its feasibility for collecting representative survey data during infectious disease epidemics remained unexplored. Roig et al. (2022) conducted a study of the gender gap related to tasks within the home during the COVID-19 health crisis, introducing the variable size of the municipality in the analyses.

Our work focuses on the combination of data obtained through probabilistic and non-probabilistic surveys with the aim of obtaining more reliable estimates through regularized logistic regression and kernel weighting techniques. As a non-probabilistic survey, we will base ourselves on the survey carried out by Pérez et al. (2022) and we will study the block of questions referring to employment and study as a probabilistic survey the CIS Barometer of May 2020.

2. Methodology

Let s_v be the non-probabilistic sample, obtained from the population of interest U, from a survey of volunteers which has a size n_v, the variable of interest y, and the vector of auxiliary variables $x = (x_1, \dots, x_p)$. In order to eliminate the volunteer bias in non-probability surveys, we must have available auxiliary information that is accurate and closely related to the topic under study. Depending on the type of auxiliary information available, we distinguish different types of bias reduction techniques (Rueda et al., 2020), although in our study we will focus on one, the kernel weighting method, due to its recent appearance and excellent results. This technique is included in the group of those that need a probabilistic reference sample in order to be carried out, from which we only need to know its auxiliary variables.

We define s_r as the probability sample, obtained from the population of interest U, which we will use as a reference and which has a size n_r, and $x = (x_1, \dots, x_p)$ as the vector of auxiliary variables that we have measured both in the reference probability sample s_r and in the non-probability (or volunteer) sample s_v. In the case of s_r, the probability that each individual has of participating in the survey is known and greater than zero, conditions that must be met for the sample to be probabilistic. This probability is known as the first-order inclusion probability (π_r). From these probabilities we obtain what are known as design weights (w_r), which will be key when forming our estimators, and which are obtained by

calculating the inverse of the π_r. In the case of non-probabilistic samples, the inclusion probabilities are not usually known since we lack a probabilistic theory to support them, making it impossible to accurately determine their value. We will therefore assume their value or, as in this case, seek to estimate these probabilities, also called propensities, for the non-probabilistic case (π_v). We define the probability of belonging to the non-probabilistic sample, $\forall i \in U$, as:

$$\pi_{vi} = P[1_i = 1|x_i] \;\; where \;\; 1_i = \begin{cases} 1 & for \;\; i \in s_v \\ 0 & otherwise \end{cases}$$

We seek to estimate the expected value of the probability of inclusion in the non-probability sample, through some model M, $\forall i \in s_v \cup s_r$:

$$\hat{\pi}_{vi} = E_M\left[\hat{1}_i = 1|x_i\right] \;\; where \;\; \hat{1}_i = \begin{cases} 1 & for \;\; i \in s_v \\ 0 & for \;\; i \in s_r \end{cases}$$

To estimate these probabilities, logistic regression models are usually considered, due to the binary nature of the variable to be estimated, although machine learning techniques can be used for the same purpose (Ferri-García and Rueda, 2020). In our work we will use what is known as regularized logistic regression, which adds penalty parameters when estimating the coefficients of the logistic regression. Other methods for obtaining these probabilities can be found at Castro-Martín et al. (2020). In the case of logistic regression, the estimated probabilities are obtained as follows:

$$\hat{\pi}_{vi} = \frac{1}{1 + exp(-\beta x_i)}, \;\; i \in s_v \cup s_r$$

being β the vector of regression coefficients. To calculate these coefficients we must maximize the log-likelihood function:

$$l(\beta) = \sum_{i \in s_v \cup s_r} \hat{1}_i \ln(\hat{\pi}_{vi}) + \left(1 - \hat{1}_i\right)\ln(1 - \hat{\pi}_{vi}) = \sum_{i \in s_v \cup s_r} \left[\hat{1}_i \ln\left(\frac{\hat{\pi}_{vi}}{1 - \hat{\pi}_{vi}}\right) + \ln(1 - \hat{\pi}_{vi})\right]$$
$$= \sum_{i \in s_v \cup s_r} \left[\hat{1}_i \beta x_i - \ln\left(1 + e^{\beta x_i}\right)\right]$$

Regularization strategies introduce penalties in order to avoid overfitting, reduce variance and minimize the influence of less relevant predictors in the model. We apply ridge regularization, which introduces an L2 penalty to the log-likelihood function. The estimated coefficients β will also be obtained by maximizing the log-likelihood function with this penalty:

$$l_R(\beta) = \sum_{i \in s_v \cup s_r} \left[\hat{1}_i \beta x_i - \ln\left(1 + e^{\beta x_i}\right)\right] - \lambda \sum_{j=1}^{p} \beta_j^2$$

The penalty depends on the parameter $\lambda \geq 0$ which measures the regularization intensity of the fit. This parameter depends on the real parameters of the regression, and since they are unknown, we will have to choose it arbitrarily or through the hyperparameter adjustment. In our work the selection of this parameter is made by means of cross-validation. This consists of dividing our data into two complementary sets, performing the analysis on one subset (training set) and validating the analysis on the other (test set).

Once we have explained how we estimate the π_v, we move on to the explanation of the technique in charge of reducing the bias, which in our work will be the kernel weighting method.

2.1. Kernel Weighting Method (KW)

Developed by Wang et al. (2020), it is based on the creation of new weights, called pseudo-weights, from the design weights of the individuals in the probability sample weighted according to the similarity they have to the non-probability sample individuals. This similarity is measured with the distance between individuals, through the difference of the estimated probabilities of belonging to the probability sample (analogous to the non-probabilistic case) and to the non-probability sample.

$$d_{ij} = \hat{\pi}_{vi} - \hat{\pi}_{rj}, \;\; i \in s_v, \;\; j \in s_r$$

These distances will have a value between -1 and 1, so we will try to smooth them. For this purpose, we make use of kernel functions centred at zero, which are continuous, symmetric, and positive functions, and can be used as density functions of statistical distributions. The closer the distance is to zero, the more similar the individuals will be with respect to their auxiliary variables, since propensities are estimated as a function of these variables. The more similar the individuals are, the more the KW will assign a higher percentage of the design weight from the individual in the probability sample to the individual in the non-probability sample. These percentages are called kernel weights, and are obtained:

$$k_{ij} = \frac{K\{d_{ij}/h\}}{\sum_{i \in s_v} K\{d_{ij}/h\}}$$

where K{.} is a kernel function centred at zero, and h is the corresponding bandwidth (Epanechnikov, 1969). The kernel weight values will be between zero and one, and the sum of all of the volunteer sample values is one. To calculate the pseudoweights KW we will sum the reference sample design weights w_r of each $j \in s_r$, weighted by the kernel weights of the i-th individual, from the non-probability sample, with each j-th individual from the reference sample. The expression of these pseudoweights is as follows:

$$w_i^{KW} = \sum_{j \in s_r} w_{rj}\, k_{ij}$$

Finally obtaining the estimator of the total:

$$\hat{Y}_{KW} = \sum_{i \in s_v} w_i^{KW}\, y_i$$

3. Application

We have applied the proposed methodology in order to correct the bias of the survey conducted during the COVID-19 lockdown in Spain carried out by Pérez et al. (2022). The survey was distributed following a snowball method via email and social networks. Therefore, a significant lack of representativity is to be expected. However, it was conducted between 28th April and 14th May, 2020, and it included some variables in common with the CIS Barometer of May 2020. The latter can be considered a reference probabilistic survey since it was carried out by an official Spanish institution following strict methodologies.

In particular, the following covariates will be used for applying the kernel weighting method: state, province, urban density, sex, age, education level, employment status, last electoral vote, intended electoral vote and confidence in the government during the pandemic.

Once we have obtained representative weights, these can be used in order to produce estimations for the economical variables included in the survey of interest. In Table 1, the differences between the estimations obtained with the naive mean and the estimations obtained after correcting the bias can be observed. Depending on the variable, the change can be quite significant.

Table 1. Naive and corrected estimations for the variables of interest (those directed to workers and students)

Variable	Naive	KW	Variable	Naive	KW
WORK.PROD	64.8	63.7	STU.PROD	86	72.8
WORK.EXP.1	17	21	STU.EXP.1	0.9	0.1
WORK.EXP.2	23.7	23.3	STU.EXP.2	7.9	8.6
WORK.EXP.3	26.2	30.2	STU.EXP.3	23.1	25.4
WORK.EXP.4	24.8	19.4	STU.EXP.4	28	26.7
WORK.EXP.5	44.9	47.2	STU.EXP.5	68	61.1
WORK.PERCEP.IN.1	27.3	32.4	STU.EXP.6	24.3	27.7
WORK.PERCEP.IN.2	5.7	8.4	STU.EXP.7	47.9	45.2
WORK.PERCEP.IN.3	10.6	11.8	STU.EXP.8	26.4	19.2
WORK.PERCEP.IN.4	3.3	1.9	STU.EXP.9	10.7	17.4
WORK.PERCEP.IN.5	59.3	55.8	STU.PERCEP.1	11.4	16.6
WORK.PERCEP.OUT.1	31.7	34	STU.PERCEP.2	49.8	46.6
WORK.PERCEP.OUT.2	8.7	8.9	STU.PERCEP.3	25	34.7
WORK.PERCEP.OUT.3	9	8.3	STU.PERCEP.4	39	45.1
WORK.PERCEP.OUT.4	55.5	53.4	STU.PERCEP.5	12.1	5.7

Note: For more information on the variables of interest, see the article on Pérez et al. (2022)

The results show the percentages of individuals which agree with each of the statements surveyed. For some general questions, such as the percentage of people who feel their work performance has been affected due to the lockdown, the results stay similar. This may be explained either by a lack of bias for those questions or because more relevant covariates are needed in order to reduce some underlying bias. In fact, when the students are asked the same question, we observe a 13.2% difference in the estimation. These changes also occur for some specific questions. For example, 5.1% more individuals than those initially observed thought that their work would be affected by an economical crisis after the pandemic.

4. Conclusions

We have evaluated a reweighting method for correcting bias in non-probabilistic surveys, by using a reference probabilistic survey. We have also confirmed with a practical application the significance of the differences in the estimations obtained. However, these differences will depend on the presence of an actual bias and on the importance of choosing the right covariates.

References

Ali, S.H., Foreman, J., Capasso, A., Jones, A.M., Tozan, Y. & DiClemente, R.J. (2020). Social media as a recruitment platform for a nationwide online survey of COVID-19 knowledge, beliefs, and practices in the United States: methodology and feasibility analysis. *BMC Medical Research Methodology.* 20, 116. https://doi.org/10.1186/s12874-020-01011-0.

Cajner T., Crane L.D., Decker R.A., Grigsby J., Hamins-Puertolas A., Hurst E., Kurz C. & Yildirmaz A. (2020). The US Labor Market during the Beginning of the Pandemic Recession. *National Bureau of Economic Research*, w27159.

Castro-Martín, L., Rueda, M. D. M., & Ferri-García, R. (2020). Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. Mathematics, 8(6), 879.

Coibion, O., Gorodnichenko Y. & Weber M. (2020). Labor Markets During the COVID-19 Crisis: A Preliminary View. *National Bureau of Economic Research*. w27017

Cowan, B.W. (2020). Short-run effects of COVID-19 on U.S. worker transitions. *National Bureau of Economic Research.* w27315. 10.3386/w27315.

De Boni, R.B. Web surveys in the time of COVID-19. (2020). *Cadernos de Saúde Pública: Reports in Public Health.* 36(7):e00155820.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications, 14(1), 153-158.

Fairlie, R.W., Couch K.A., & Xu H. (2020) The Impacts of COVID-19 on Minority Unemployment: First Evidence from April 2020 CPS Microdata. *National Bureau of Economic Research*, w27246.

Ferri-García, R., & Rueda, M. D. M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PloS one, 15(4), e0231500.

Gourlay, S., Kilic, T., Martuscelli, A., Wollburg, P., & Zezza, A. (2021). High-frequency phone surveys on COVID-19: good practices, open questions. Food Policy, 105, 102153.

Hlatshwako, T.G., Shah, S. J., Kosana, P., Adebayo, E., Hendriks, J., Larsson, E.C., Hensel, D.J., Erausquin, J.T., Marks, M., Michielsen, K., Saltis, H., Francis, J.M ., Wouters, E., & Tucker, J.D. (2021). Online health survey research during COVID-19. *The Lancet: Digital Healt.* 3. https://doi.org/10.1016/S2589-7500(21)00002-9.

Peréz, V., Aybar, C. & Pavía, J.M. (2022). Dataset of the COVID-19 lockdown survey conducted by GIPEyOP in Spain. *Data in Brief*, 40, https://doi.org/10.1016/j.dib.2021.107700.

Roig, R., Aybar, C., & Pavía, J. M. (2022). COVID-19, gender housework division and municipality size in Spain. Social Sciences, 11(2), 37.

Rueda, M.D.M., Ferri-García, R., & Castro, L. (2020). The R package Non-ProbEst for estimation in non-probability surveys. The R Journal, 12(1), 406-418.

Schieman, S., Badawy, P.J., Milkie, M.A. & Bierman A. (2021). Work-life conflict during the COVID-19 pandemic. *Socius: Sociological Research for a Dynamic World*, 7, 10.1177/2378023120982856.

Wang, L., Graubard, B. I., Katki, H. A., & Li, A. Y. (2020). Improving external validity of epidemiologic cohort analyses: a kernel weighting approach. Journal of the Royal Statistical Society: Series A (Statistics in Society), 183(3), 1293-1311.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16041

Finding patterns from a user-centric perspective using knowledge discovery methods

Arturo Palomino[1,2], Karina Gibert[2]
[1]Lidl, [2]Intelligent Data Science and Artificial Intelligence Research Center, Univesitat Politècnica de Catalunya, Barcelona, Spain

Abstract

Chained advertisement involves breaking down a marketing campaign message into multiple banners that are shown to a user in a specific sequence in order to create a less intrusive and more effective campaign. The challenge is determining the most effective sequence of websites and banner order. This study aims to develop a recommendation system to assist with this issue. To address the vast size of the internet and the complexity of the problem, the research uses a data-driven computational approach to estimate the probability of different sequence events and apply this to real user data from a leading company. The proposed method is faster and more efficient than previous approaches.

Keywords: *user-centric clickstream; sequence; profiling; chained advertisement; recommender systems; probability.*

1. Introduction

From the early 90s with the introduction of WWW (World Wide Web) protocol by Tim Berners-Lee (Berners-Lee et al. 2004), Internet has revolutionized industries and has become an opportunity for manufacturers, media agencies, market research companies. Media analysis traditionally focuses on three issues: optimization and Return of Investment (ROI) (Powell, 2012), AB-test (Hayes, 2006) and Marketing Mix Models (MMM) (Borden, 1964).

Online has been introduced as a new advertising channel (Xu, 2014) (Young, 2014), leading to the modification of MMM models to include a factor that reflects the impact of online marketing campaigns on consumer purchases, together with TV, Press, and other effects. AB-Test (and Pre-Post analysis) compares two samples using a classical case-control scheme: one impacted by a campaign and a control group not exposed to advertisements. The campaign's effect is measured, comparing sales of both groups.

Internet offers more detailed information about users than simply whether or not they have seen an advertisement, and current interests in marketing research focus on using this information to improve predictive models and campaign designs. Chained advertisement, which involves dividing a marketing campaign message into parts and presenting them in a sequence of banners. It is particularly suitable for the internet and results in less intrusive advertising that can guide users towards a possible online purchase. Chained advertisement requires determining the proper sequence of websites to place banners in order to maximize the campaign's impact. This study proposes a recommender system that uses user browsing information to identify the most visited routes on the internet and determine the most effective websites for banner placement. The system uses a data mining methodology considering the order in which domains are visited and doesn’t assume the Markov premise. It is tested with a real data and designed to be scalable in Big Data scenario.

The structure of the paper is the following. In section 1 this research is introduced and motivated. In section 2 related work is described. In section 3, the formal structure of the problem is presented, and section 4 presents the joint probability discussion. In section 5 the methodology is described. In section 6 general and conditional solutions are explained. In section 7 the use of the proposal to make the targeted recommendations to design a chained advertisement publicity is shown. In section 8 conclusions and future work are discussed.

2. Related work

As said before, the information about user’s browsing activities is provided by clickstream data and recorded in log files. One of the first references discussing about the opportunities and effects associated to clickstream data analysis is (Florey, 1996). In (MacDonald, 1999) clickstream data is classified in three big groups, according to the collection scheme used:

There are three types of data collection methods: Site-centric, Ad-centric, and User-centric. Each approach has different level of detail, User-centric method provides more detailed information about browsing activity. Currently, the most popular approach is the site-centric data. Indeed, there are commercial products suitable to analyze log files containing site-centric data, like Google Analytics and Adobe Analytics, and literature is abundant. Under this approach, only previous and own visited web is available, most of the works assume Markov hypothesis to predict permanence time or revisits. Other works use clustering methods for segmentation of web users, modeling, and forecasting (Wang, 2004) (Pei et al. 2001). But in major part of the reference models are limited to predict churn or repetition, Next Visit or Last click. Other works use site-centric data to build Recommender systems (Adomavicius et al. 2001; Van den Poel et al. 2004) based on the contents use and similar user's identification. Fewer references are found on Ad-centric data, some based on the use of cookies to collect information (Moe, 2003).

Site-centric and Ad-centric data collection has limitations for understanding complete user browsing activity and linking it to sociodemographic characteristics in chained advertisement contexts. User-centric data collection addresses these limitations but is costly. Most studies use this data for proactive recommendations and characterizing visited sites without considering the visits order. The order is crucial for chained advertisement and maximizing the probability of the user receiving the pieces of publicity in the intended order.

User-centric approach needs representative panels and is costly and difficult to maintain. Indeed, 4 big companies can be found providing clickstream data: Nielsen's CDR, Gfk's Netquest (Revilla, 2017), Alexa Internet (Vaughan, 2012) and ComScore's Mediametrix. User-centric data is underutilized, only used for exposure to the banner and not for final purchase analysis, despite the availability of complete browsing and sociodemographic information. Even a simple analysis of this data can be helpful for chained publicity, such as identifying the most visited sequence of sites of a given population, to identify the websites where the pieces of communication should be placed for optimal ROI.

Authors are not aware of works finding the most frequent sequence of webs visited by users, but in the field of sequence pattern mining, some proposals are found to identify sequences (Balcázar et al. 2007; Srikant, 1996; Han et al. 2000). Even though most works are not related to marketing or clickstreams, they have been analyzed for potential application in clickstream data analysis. Three families of methodologies are identified:

• Apriori like methods (Agrawal et al. 1995): like GSP (Srikant, 1996) and AprioriAll (Agrawal et al. 1996). Frequent itemsets are used to filter irrelevant information for efficiency purposes.

• Pattern grow: like FreeSpan (Han et al. 2000) and PrefixSpan (Han et al. 2001). The data base is filtered while iterating on what is called a projection of data base, where only baskets starting with the sequence of last step are taken into consideration.

• Vertical format of database: Data is preprocessed in a specific order to efficiently extract the most frequent sequences. Best examples are SPAM (Ayres, 2002) and SPADE.

Previous methods for identifying patterns in data variability is allowed in the pattern and do not require elements to be in contiguous form. For the case of browsing, these methods will provide patterns consisting of a set of sites visited one after the other, but in between each two, the user might have jumped to other sites. This conception is not much suitable for the context of chained advertisement, where showing all the pieces of the message in the right sequence and without external interruptions is crucial for the impact of the campaign. Therefore, new methodological approaches become necessary to allow the analysis of user-centric clickstream data for both identifying patterns of ordered and contiguous sequences of sites and quantifying their associated probability without making Markov assumptions. The novelty of the method presented in this paper is that it finds sequences with adjacent and sorted sites instead of using unsorted and non-contiguous bags of items.

3. Formalization

Our aim is to find the most likely sequence of sites visited by users using joint probability distribution. Given a set of internet users I = $\{\boldsymbol{i1}\ldots, \boldsymbol{in}\}$ browsing on the network; The space of Internet domains available $\mathbb{D}$ = {k}$\forall \boldsymbol{k} \in \mathbb{N}+$. The space of all possible routes that a user can follow in an Internet session is: $\boldsymbol{\mathcal{R}} = \boldsymbol{D} \cup (\boldsymbol{D} \times \boldsymbol{D}) \cup (\boldsymbol{D} \times \boldsymbol{D} \times \boldsymbol{D}) \cup \ldots = \bigcup_{j=1}^{\infty} \boldsymbol{D}^j = P_{\mathcal{R}}$

$\boldsymbol{r} \in \boldsymbol{\mathcal{R}}$ represents an internet walk of a given user during a session. $\forall \boldsymbol{r}\, \exists \boldsymbol{l}$ so that $\boldsymbol{r} \in \boldsymbol{D}^l$ where $\boldsymbol{l} \in \mathbb{N}^+$ is the length of the route. $\boldsymbol{r}$ is expressed as a limited sequence of domains $\boldsymbol{r} = \{\boldsymbol{d_1}, \ldots, \boldsymbol{d_l}\}$. Given $P_{\mathcal{R}}$, the probability law associated to $\mathcal{R}$ so that $\forall \boldsymbol{r} \in \mathcal{R}$, $P_{\mathcal{R}}(\boldsymbol{r})$ is the probability of a user following route $\boldsymbol{r}$ in a session, the underlying probability problem to be solved is to maximize the probability function of $\mathcal{R}$, by identifying r such that:

$$\boldsymbol{r} \in \mathcal{R}: P_{\mathcal{R}}(\boldsymbol{r}) = \max_{\forall s \in \mathcal{R}} P_{\mathcal{R}}(s)$$

4. Finding the joint probability distribution of a sequence of events

To quantify the probability of each element in $\mathcal{R}$, which is a huge events space $\boldsymbol{\mathcal{R}}$, computational statistics should help to find the maximum route. Let $\boldsymbol{S_p}$ $(\boldsymbol{p} \in \mathbb{N}^+)$ be the domain visited in $\boldsymbol{p}$-th position of the session. $\boldsymbol{S_p}$ can take values from $\boldsymbol{D}$, $\boldsymbol{S_p} = \boldsymbol{D}$. The probability of r is:

$$P_{\mathcal{R}}(\boldsymbol{r}) = P_{\mathcal{R}}(d_1, d_2, \ldots, d_l) = P(S_1 = d_1, S_2 = d_2, \ldots, S_l = d_l) \quad \forall\, l \in \mathbb{N}^+$$

It is usual to use Markov assumption to compute joint probabilities, i.e., $\boldsymbol{p}$ only depends on the domain $\boldsymbol{p}$**-1.** Being $P_{p_{k_2 k_1}}$, the probability of visiting d_{k_2} at $\boldsymbol{p}$, from d_{k_1} in position $\boldsymbol{p}$-1, is:

$$P_{p_{k_2 k_1}} = P(\boldsymbol{S_p} = d_{k_2} | \boldsymbol{S_{p-1}} = d_{k_1}),\ \forall k_1, k_2 \in \boldsymbol{D}$$

In a scenario where Markov assumption holds:

$$P_{k_1,k_2..,k_l} = P(\boldsymbol{S_1} = d_{k_1}) \prod_{p=1}^{l} P\left(S_p = d_{k_p} | S_{p-1} = d_{k_{p-1}}\right)$$

Thus, joint probability can be calculated in terms of the conditional probabilities of arriving to a certain web domain, given the previous one. Figure 1 displays the transitions between previous and posterior domain of the walk at a given position $\boldsymbol{p}$. However, the internet walks of users have memory, and Markov cannot be used. Considering r1: google→facebook→live→youtube. The proposed algorithm finds in efficient time the probability of this sequence P(r1)= 0,006. Assuming Markov property: P(r1)= P(youtube|live) P(live)= 0,0000579. Even more, the independence assumption between domains is also non holding: P(r1)= P(youtube|live) P(live|facebook) P(facebook|google) P(google)= 0,000002. Whereas P(r1)= P(youtube |live, facebook, google) P(live |facebook, google) P(facebook |google) P(google)= 0,006, as expected. In this research, independence and Markov properties will not be assumed, so that:

$$P_{k_1,k_2..,k_l} = \sum_{\forall d_{k_1}} \sum_{\forall d_{k_2}} \dots \sum_{\forall d_l} P(S_l) = d_l | S_{l-1} = d_{l-1}, S_{l-2} = d_{l-2}, \dots, S_1 = d_1) \dots \cdot P(S_p = d_p | S_{p-1} = d_{p-1}, S_{p-2} = d_{p-2,\dots} S_1 = d_1) \dots \cdot P(S_2 = d_2 | S_1 = d_1) \cdot P(S_1 = d_1)$$

Considering that card(D)=280 million domains, building this probability function is still unaffordable. In fact, with the complete WWW universe, the number of potential routes which could be followed by a user is: $\text{card}(\mathcal{R}) = \sum_{l=1}^{\infty} card(\boldsymbol{D^l}) = \sum_{l=1}^{\infty} (280E10^6)^l$, which is huge. Just to have an idea, the first 20 terms of this series make a total of 8,7733E+168 potential $\boldsymbol{l} \leq 20$ routes

The probability of a certain page can be computed using the whole sequence of previous pages without assuming Markov assumption. The computation of the joint probability function of routes cannot be reduced to simple conditioning of immediately previous domain, neither to the simple product of marginal domains. Surfer's interests during navigation follow an objective that guide the sites visited. It cannot be assumed, for instance, that probability of visiting **zara.com** is the same coming from **dior.com** and then **mango.com** than if we came from **berska.com** and then **mango.com**.

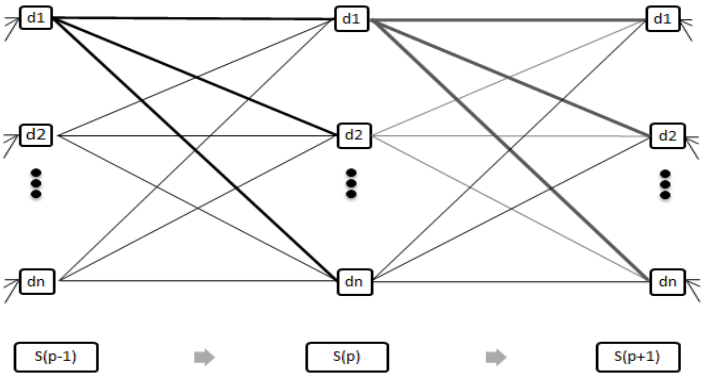


Figure 1. Domain transitions

The availability of user-centric data, with the whole user route, is more predictive than site-centric or add-centric data. The latter can only provide the immediately previous domain visited, assuming memory loss. Thus, the proposed frequentist approach provides more reliable estimates of the probabilities of a given route. A computational approach estimates the joint probability function for a particular sample of users in a certain context. The main idea in behind is simple:

Given a route $\boldsymbol{r} \in \mathscr{R}$: $P_{\mathcal{R}}(\boldsymbol{r}) = \lim_{n\to\infty} \frac{n_r}{n}(s)$, where n_r is the number of occurrences of r_i in a sample and n users, according to the frequentist principle of probability. As we are in a Big data frame, the sample size n is big enough to guarantee convergence. But not all $\boldsymbol{r} \in \mathscr{R}$ are to be considered. Reducing to the observed set of occurring sequences in a given sample, sensibly limits the dimension of the event's space. In particular, for a real sample in a certain month, the number of different domains visited by the users is about 40000, and considering walks no longer than 20 websites, the total number of potential routes reduces about 1,09954E+92. Considering a complete year data, the number of different domains visited by the users is about 300000, and the total number of potential routes (no longer to 20 websites) moves to 3,4868E+109, in any case extremely lower than card($\mathcal{R}$). Therefore, our approach will be to estimate all probabilities not in an analytical way but with a computational process that will be able to extract all observed routes in a simple screening of the database. From a computational point of view computing n_r is extremely time consuming when faced by means of brute force algorithm.

$$n_r = n_{\mathrm{k}_1,\mathrm{k}_2..,\mathrm{k}_l} = card\{i \in I : \boldsymbol{r_i} = (\boldsymbol{d}_{\mathrm{k}_1}, \dots, \boldsymbol{d}_{\mathrm{k}_l}), \boldsymbol{r_i} \in \mathcal{R}\}$$

Thus, our proposal is to compute the observed frequencies of each route as a proportion of occurrences in the database to estimate the function *p*. Eventually, this work is user-centric, and we have access to users' sociodemographic information. In this paper we will also focus on the most frequent routes taken by a certain demographic profile, for marketing purposes. This corresponds to solve a variation of the original problem, being A the specific profile targeted:

$$\boldsymbol{r} \in \mathcal{R} : P_{\mathcal{R}|A}(\boldsymbol{r}) = \max_{\forall s \in \mathcal{R}|A} P_{\mathcal{R}|A}(s)$$

5. Methodology

At this point the problem has been reduced to count how many times each sequence appears in the database. The space of sequences is huge and brute force is expensive. Reducing only to observed sequences is relevant. We have developed a patented procedure (Palomino et al. 2023) which is able to find the sequences of a given length $\boldsymbol{l}$ and quantify the empirical joint distribution function in highly scalable conditions. Table 1 shows the CPU time for both the sample of one week data and one year data, for the identification of all sequences of length $\boldsymbol{l}$ =4. The initial logs are large (310,785.233 registers), include information about CSS files, Jscript, DoubleClick, tags, chats, agents, etc, but the number of rows containing information about the domains voluntarily visited by users is smaller (8,008.565) but still important. Whereas the ratio between useful rows analyzed between one year data and one week data is 52:1 the ratio of CPU times is 6:1, which indicates a less tan linear trend. It can synthetize a database of 8 million rows in most frequent sequences in only 2 minutes. Moreover, regarding CPU time obtained in the previous proposals (Palomino et al. 2014) for sequences of $\boldsymbol{l}$ =4, the current proposal (Palomino et al. 2023) is sensibly reducing CPU time (Table 2).

Table 1. Time elapse between one week and one year samples

	1 Year (L=4)	1 Week (L=4)
Initial size log	310.785.223	6.678.800
Useful rows	8.008.565	156.954
CPU time	120 ‘’	18 ‘’

Table 2. Time elapse between one week and one year samples.

Algorithm	CPU time (1 Week data, L=4)
Brute Force	3:05:00
Apriori-based	00:54:00
Tabulation-based	00:26:00
New proposal	00:00:18

6. Recommending sequences of sites for chained marketing campaigns

In this work, real data provided by the operational company Compete (WPP's company) is analyzed. Data comes from a continuous panel of internet browsing habits, representative of the 12 million of internet Spanish households. Data gathers clickstream user-centric and sociodemographic information (details in (Palomino et al. 2014) and (Palomino et al. 2018)). A large retail company wants to launch a new product of premium high quality sport shoes for babies of less than 5 years with a chained advertisement composed by 3 banners. The marketing leader is interested in the following target population:

- Madrid household, young couples with children, both younger than 50 years, high social class.

The goal is to identify and quantify the sequences of websites more frequently visited by the target profile of users (Table 7). The Top sequence according to the number of households is: **google.com→marca.com→williamhill.com.** With this information, the recommendation is to advertise first banner on google, second on marca, and third on williamhill (Figure 2).

Table 7. Frequent sequences for specific target

domini 1	domini 2	domini 3	freq	Llars
google.com	**marca.com**	**williamhill.com**	62	19
google.com	facebook.com	hotmail.com	21	14
live.com	facebook.com	google.com	21	13
google.com	**marca.com**	facebook.com	79	12
google.com	live.com	facebook.com	22	11

It's important to note that these figures are from a panel sample, must be combined with scaling factors to estimate the total population represented by these households, as is standard

in continuous panel methodologies. This is not covered in this work, but it's included in the subsequent part of the recommender.

7. Conclusions and further work

In this paper we use clickstream data with socio-demographic information to create a marketing campaign recommender system. The approach identifies the optimal sequence of domains to place de sequence of banners of a chained publicity campaign. This work aims to understand and formalize the problem of finding all routes, probabilities, and rankings of web domains visited by a sample population using an empirical, data-driven approach, avoiding the use of Markov assumption and the "bag of sites" approach used in other works for different contexts. An efficient algorithm has been developed and implemented to identify all sequences followed by users (Palomino et al. 2023) that supposes an improvement over previous development (Palomino et al. 2014).

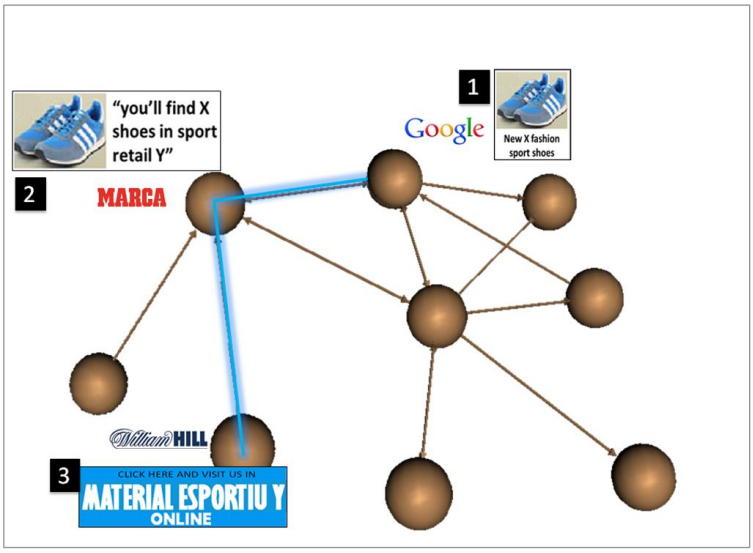


Figure 2. Targeted campaign for a specific product.

Moreover, this work adds to the support (number of sessions following the patter) a second optimality indicator, the number of households. In this research, a large database from a leading company, Compete, is used to confirm in a real case, with the implementation of the algorithm, that the CPU computation time is sensibly improved and highly scalable, and can be used for targeting any kind of subpopulations. In a further analysis more ambition goals are considered, like the addition of new optimality indicators, and the introduction of the cost of the campaign to enrich recommendations with ROI assessment.

References

Adomavicius, G., & Tuzhilin, A. (2001). Expert-driven validation of rule-based user models in personalization applications. DM&KD, 5(1-2), 33-58.

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In Data Engineering, 1995. Procs of the 11th Int'l Conf IEEE (pp. 3-14).

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. Advances in knowledge discovery and data mining, 12(1), 307-328.

Ayres, J., Flannick, J., Gehrke, J., & Yiu, T. (2002). Sequential pattern mining using a bitmap representation. In Procs 8th ACM SIGKDD 429-435

Balcázar, J. L., & Garriga, G. C. (2007). Horn axiomatizations for sequential data. Theoretical Computer Science, 371(3), 247-264.

Borden, Neil H. "The concept of the marketing mix." Journal of advertising research 4.2 (1964): 2-7.

Florey, K., (1996) Who's Been Peeking At My Clickstream?. Ethics and Law on the Electronic Frontier. MIT, Cambridge, MA, 1995

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., & Hsu, M. C. (2000, August). FreeSpan: frequent pattern-projected sequential pattern mining. En Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. 2000. p. 355-359.

Han, J., Pei, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. (2001) Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. En proceedings of the 17th international conference on data engineering. IEEE, p. 215-224.

Hayes, G. (2006). Social Cross Media – What Audiences Want. Retrieved 02 26,. 201

MacDonald, C.S. (1999), Evolving models of online audience measurement: Developments since Vancouver. Worldwide Readership Research Symposium (1999) 9.1 487-492

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. Journal of consumer psychology, 13(1), 29-39

Palomino A., & Gibert, K. (2014). Web pattern detection for bussiness intelligence with data mining. FAIA 269: 277-280 IOSPress, NL.

Palomino, A., & Gibert, K. (2018). Web Pattern Navigation Profiling for Online Marketing Campaigns Design Support Under a Data Science Approach. En CCIA. 2018. p. 166-175.

Palomino A., & Gibert, K. (2023). Solicitud de Patente en España Nº P202330024; solicitante UPC.

Powell, G. R. (2012). Marketing Calculator: Measuring and Managing Return on Marketing Investment. John Wiley & Sons.

Revilla, M., Ochoa, C., & Loewe, G.(2017). Using passive data from a meter to complement survey data in order to study online behavior. Social Science Computer Review, vol. 35, no 4, p. 521-536.

Srikant, R., & Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements (pp. 1-17). Springer Berlin Heidelberg. ISO 690

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. European Journal of Operational Research, 166(2), 557-575.

Vaughan, L. (2012) An alternative data source for web hyperlink analysis:"sites linking in" at Alexa Internet. Collnet journal of scientometrics and information management, 2012, vol. 6, no 1, p. 31.

Wang, J., & Han, J. (2004). BIDE: Efficient mining of frequent closed sequences. In Data Engineering. Proceedings. 20th International Conference on (pp. 79-90). IEEE.

Xu, J., Forman, C., Kim, J. B., & Van Ittersum, K. (2014). News media channels: complements or substitutes? Evidence from mobile phone usage. Journal of Marketing, 78(4), 97-112.

Young, A. (2014). Google and Facebook. In Brand Media Strategy (pp. 7-14). Palgrave Macmillan US.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16446

Productivity, digital footprint and sustainability in the textile and clothing industry

Josep Domenech[1], Ana Garcia-Bernabeu[1] and Pablo Diaz-Garcia[2]
[1]Department of Economics and Social Sciences, Universitat Politècnica de València Spain,
[2]Department of Textile and Paper Engineering, Universitat Politècnica de València, Spain.

Abstract

In recent years, there has been a shift from the linear economic model on which the textile and clothing industry is based to a more sustainable model. However, to date, limited research on the relationship between sustainability commitment and firm productivity has focused on the textile and clothing industry. This study addresses this gap and aims to explore whether the digital footprint of small and medium-sized textile companies in terms of their sustainable performance is related to their productivity. To this end, the paper proposes an innovative model to monitor the companies' commitment to sustainable issues by analyzing online data retrieved from their corporate websites. This information is merged with balance sheet data to examine the impact of sustainability practices, capital and human capital on productivity. The estimated firm's total factor productivity is explained as a function of the sustainability digital footprint measures and additional control variables for a sample of 315 textile firms located in the region of Comunidad Valenciana, Spain.

Keywords: *Productivity; Digital footprint; Web scraping; Sustainability assessment; Textile and clothing industry.*

1. Introduction

Sustainability has gained growing attention in the textile and clothing (T&C) industry as it has turned into one of the most polluting industries in the world. After the oil industry, it is responsible for 10% of carbon emissions (Manshoven, et al. 2019). In Europe, apparel, footwear, and home textiles are the fourth most pressured sector in terms of primary resource and water use, after food, housing, and transport (Ellen MacArthur Foundation, 2017). According to Euratex, in 2021, the entire EU-27 T&C industry represented a turnover of € 147 billion and 143000 companies, mainly micro and SMEs (EURATEX, 2022). In this context, the textile and clothing industry has started to develop forward-looking business models to reconcile competitiveness with sustainability.

The idea that adopting sustainable practices leads to improved corporate performance is well supported by the literature. The link between sustainability and the economic performance of firms has been investigated from theoretical and applied points of view. Russo and Fouts (1997) argued that improvements in environmental performance lead to competitive advantages in terms of cost reduction, enhanced reputation, and increased competitiveness. Focusing on Spanish SMEs, Jorge et al. (2015) suggest that environmental performance has a positive impact on competitive performance as well as the mediating effects of image and relational marketing. Other studies (Aragon-Correa et al., 2008; Galdeano-Gomez et al., 2008) also confirmed the existence of a direct and positive relationship between financial performance and environmental strategies.

A critical debate in the literature concerns the measurement of the sustainability at the corporate level (Pranugrahaning et al., 2021), particularly in SMEs, whose reporting practices have been scarcely researched (Martins et al., 2022). In the lack of a standardized framework for assessing corporate sustainability, many companies, including SMEs, have chosen to report their engagement in sustainable practices through their corporate website (Palma et al. 2022; Lodhia, 2010).

The purpose of this paper is to analyze how textile firms' commitment to sustainability, measured through web content analysis, impacts productivity. The research addresses two main questions: Is it possible to measure the commitment to sustainability of textile companies through their digital footprint? Has the sustainability commitment of textile and apparel firms a positive effect on productivity performance? The research measures the intensity of firms' sustainability commitment by applying web scraping techniques to their corporate websites. The impact on productivity is measured by means of the Total Factor Productivity (TFP), estimated using Levinsohn and Petrin (2003) estimator. To the best of our knowledge, this is the first time that the relation between sustainability commitment, measured with web scraping techniques, and productivity is studied.

2. Corporate sustainability assessment

The demand for sustainable development is challenging for companies both socially and institutionally. In recent years, there has been significant progress in aligning business objectives with the Sustainable Development Goals (United Nations and Development, 2015) since the UN established the 2030 Agenda in 2015. This has led to a growing interest in the academic community on corporate sustainability assessment as a new field of research, which provides tools that can support businesses transitioning towards sustainable development (Pranugrahaning et al., 2021).

There are two major approaches to assessing sustainability at the corporate level: norms and standards and single and composite indicators. Sustainable norms and standards include the Global Reporting Initiative (GRI), the OECD Sustainable Manufacturing Toolkit, and the ISO 14001 Environmental Management Systems (EMS). The GRI produces a comprehensive Sustainability Reporting Framework, which sets out the principles and indicators that organizations can use to measure and report their economic, environmental, and social performance. The OECD Sustainable Manufacturing Toolkit includes an internationally applicable common set of indicators to help businesses measure their environmental performance at the level of a plant or facility. ISO 14001 EMS is considered the leading management tool for addressing environmental degradation. Companies that become certified as complying with these ISO standards by third-party audit can demonstrate their commitment to sustainability by monitoring, managing, and improving their environmental performance. ISO 50001 is the most widely used corporate energy management standard in the world.

In addition to these standards, public and private organizations are increasingly promoting voluntary and non-profit initiatives in favor of sustainability, particularly within the textile industry. The Better Cotton Initiative (BCI), the Recycled Claim Standard (RCS), the Organic Content Standard (OCS), and the Global Recycled Standard (GRS) are some of the most relevant initiatives for promoting sustainability in the textile industry.

Single and composite corporate sustainability indicators are another approach to assessing sustainability. However, measuring corporate sustainability is multidimensional, and there is no clear consensus on which set of indicators to use to manage and measure corporate sustainability performance (Montiel and Delgado-Ceballos, 2014). Despite international recommendations, every different case in the academic literature states its own criteria and indicators, which leads to confusion among practitioners (Buyukozkan and Karabulut, 2018). Furthermore, the need for practical guidance to measure and sustainability performance has been suggested by Moldavska and Welo (2019).

Studies on sustainability performance indicators for the textile industry are recent. Ren (2000) suggested a methodology for developing sector-specific environmental performance

indicators for textile processes and products. According to Luo et al. (2021), the measurement of sustainability performance in the textile and apparel industry can be categorized into four main methodologies: life cycle assessment, environmental footprint, eco-efficiency, and the Higg-Index. The Higg-Index, developed by Sustainable Apparel Coalition (SAC), is a comprehensive set of ratings to track and measure the environmental and social impact of apparel and footwear products and companies on a scale of 0 to 100.

3. Company websites as a source of information

Company websites are a valuable source of information that reflects a company's behavior and identity. Analyzing website content can provide insights into a company's market orientation, innovative behavior, and survival rates (Axenbeck and Breithaupt, 2021; Heroux-Vaillancourt et al., 2020; Blazquez et al., 2018). Technological advancements have enabled the automation of website analysis through web crawling and web scraping techniques, which can extract data from websites and convert it into structured data suitable for analysis (Kumar et al., 2017; Diouf et al., 2019).

The benefits of using company websites as a research tool include their public accessibility, convenience, objectivity, granular data, and up-to-date information (Gok et al., 2015; Blazquez and Domenech, 2018; Hillen, 2019). However, limitations include potential biased or incomplete information, limited data availability due to anti-bot techniques, or legal and ethical concerns (Basso and Sicco, 2009; Krotov and Johnson, 2022; Luscombe et al., 2022).

Despite the limitations, the use of company websites as a research tool remains advantageous, especially for analyzing specific companies in detail. The availability of data also allows for the tracking of changes in company behavior over time. Researchers should be aware of the limitations and take measures to ensure the accuracy and ethics of the data collected through web scraping techniques. Overall, the use of company websites as a research tool is a valuable and convenient method for gathering insights into a company's activities, intentions, and strategies.

4. Methods

4.1. Empirical model

The Cobb-Douglas production function is a widely used theoretical framework in the productivity literature to describe the relationship between factors of production and output in a production process. This production function takes the form:

$$Y_{it} = A_{it} K_{it}^{\beta_k} L_{it}^{\beta_l} \quad (1)$$

where Y is the output of the production process of firm i in period t, A is the total factor productivity, K, and L are inputs of capital and labor, respectively. This function is easily linearized by taking natural logarithms. The estimation of the production function is affected by simultaneity bias due to the fact that productivity is not directly observable. The Levinsohn and Petrin (2003) estimator addresses this issue by expressing the unobserved productivity as a function of observable variables such as intermediate materials and capital stock. Once Equation (1) is estimated, the productivity analysis involves the regression of TFP, measured as $log(A_{it})$, on various web measures of corporate sustainability and additional control variables:

$$TFP_{it} = \alpha_t + \gamma_S S_i + \gamma_X X_{it} + u_{it} \quad (2)$$

where α_t are time-specific effects, S_i is a vector of variables related to corporate sustainability, and X_{it} is a vector of control variables with firms' characteristics affecting their productivity level. Sustainability was measured by counting the number of different sustainability-related words that were found on the website. Three word lists were considered: i) an extensive list of general and specific concepts (*nkeywords*), ii) a short list of general concepts (*nkwgeneral*), and iii) a list of certifications related to sustainability (*ncertif*). The control variables were chosen based on the productivity literature and included export orientation, firm age, and the gender of the manager.

4.2. Data

The sample for this study covers 315 textile firms located in the Comunidad Valenciana region in Spain, with data for the years 2020 and 2021 considered. The productivity and control variables were retrieved from the SABI database. As for the sustainability variables they were extracted from the websites of the companies after crawling the complete website.

5. Results

Equation (1) was estimated using the *prodest* R-package (Rovigatti, 2017), and its results were employed to estimate the TFP at the firm level. Four different specifications were considered to examine the association of sustainability reporting with a company's productivity. The estimation results are presented in Table 1.

Model I includes *nkeywords* as a measure of sustainability intensity. The results suggest that the TFP increases by 1.1% for each additional sustainability-related keyword found on the company's website. This remains robust, even after controlling for other variables in the regression, as Model III indicates.

Model II distinguishes between two categories of sustainability-related terminology: broad concepts (*nkwgeneral*) and certification-related (*ncertif*). The estimation results indicate that

the use of certification-related words has a more substantial impact on TFP compared to broad sustainability terms. Specifically, the presence of certification-related words is associated with a 16% increase in TFP, while the presence of each different broad sustainability term is associated with a 4% increase in TFP. These effects are slightly more pronounced when control variables are included in the equation (Model IV).

Table 1. Effect of sustainability on productivity

	Model I	**Model II**	**Model III**	**Model IV**
nkeywords	0.011** (0.003)		0.012** (0.003)	
nkwgeneral		0.040** (0.012)		0.041** (0.013)
ncertif		0.160* (0.054)		0.174* (0.053)
Export			0.011 (0.041)	0.010 (0.041)
Woman			–0.057 (0.057)	-0.055 (0.057)
log(Age)			–0.132*** (0.043)	-0.135*** (0.042)
(Constant)	1.507 *** (0.036)	1.493*** (0.036)	1.938*** (0.148)	1.937*** (0.149)
N	608	608	587	587
R^2	0.025	0.041	0.043	0.061

Dependent variable: TFP. Robust standard errors in parentheses. Time-specific effects included.
*** p < 0.001; ** p < 0.01; * p < 0.05

6. Conclusions

This study provides some evidence about the impact of adopting sustainable practices on the productivity of T&C companies. One of the key contributions of this paper is the way in which the firm's intensity of sustainability is measured by analyzing online data retrieved from their corporate websites. Secondly, the estimation of TFP uses the Levinsohn and Petrin methodology. Next, we related TFP with the intensity of firm'sustainability and additional control variables.

Our findings on a sample of 315 small and medium-sized enterprises of the T&C industry located in the Comunidad Valenciana confirm that the association of different measures of sustainability reporting with firms' total factor productivity is positive and significant. Results are robust after controlling for other variables. The main limitation of the study is that the sample refers only to companies in the Comunidad Valeniana region, therefore, as a

future line of research, it is proposed to extend the sample to companies in the national and international context.

Acknowledgments

This work was partially funded by MCIN/AEI/10.13039/501100011033 under grant PID2019-107765RB-I00.

References

Aragon-Correa, J. A., Hurtado-Torres, N., Sharma, S., and García-Morales, V. J. (2008). Environmental strategy and performance in small firms: A resource-based perspective. *Journal of environmental management*, *86*(1), 88–103.

Axenbeck, J. and Breithaupt, P. (2021). Innovation indicators based on firm websites—which website characteristics predict firm-level innovation activity? *PloS one*, *16*(4), e0249583.

Basso, A. and Sicco, S. (2009). Preventing massive automated access to web resources. *Computers and Security*, *28*(3-4), 174 – 188.

Blazquez, D. and Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99–113.

Blazquez, D., Domenech, J., and Debon, A. (2018). Do corporate websites' changes reflect firms' survival? *Online Information Review*, *42*(6), 956–970.

Buyukozkan, G. and Karabulut, Y. (2018). Sustainability performance evaluation: Literature review and future directions. *Journal of environmental management*, *217*.

Ciliberti, F., Pontrandolfo, P., and Scozzi, B. (2008). Investigating corporate social responsibility in supply chains: a sme perspective. *Journal of cleaner production*, *16*(15), 1579–1588.

Diouf, R., Sarr, E. N., Sall, O., Birregah, B., Bousso, M., and Mbaye, S. N. (2019). Web scraping: state-of-the-art and areas of application. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 6040–6042. IEEE.

Ellen MacArthur Foundation (2017). A new textiles economy: redesigning fashion's future. *Ellen MacArthur Foundation*, pages 1–150.

EURATEX (2022). Facts and key figures of the textile and clothing industry. https://euratex.eu/wp-content/uploads/EURATEX FactsKey_Figures_2022rev-1.pdf

Galdeano-Gomez, E., Cespedes-Lorente, J., and Martinez-del Rio, J. (2008). Environmental performance and spillover effects on productivity: evidence from horticultural firms. *Journal of environmental management*, *88*(4), 1552– 1561.

Gok, A., Waterworth, A., and Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*, 653 – 671.

Heroux-Vaillancourt, M., Beaudry, C., and Rietsch, C. (2020). Using web content analysis to create innovation indicators—what do we really measure? *Quantitative Science Studies*, *1*(4), 1601–1637.

Jorge, M. L., Madueño, J. H., Martinez-Martinez, D., and Sancho, M. P. L. (2015). Competitiveness and environmental performance in Spanish small and medium enterprises: is there a direct link? *Journal of cleaner production*, *101*, 26–37.

Krotov, V. and Johnson, L. (2022). Big web data: Challenges related to data, technology, legality, and ethics. *Business Horizons*.

Kumar, M., Bhatia, R., and Rattan, D. (2017). A survey of web crawlers for information retrieval. *WIREs Data Mining Knowl Discov, 7: e1218*.

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*, *70*, 317– 341.

Lodhia, S. K. (2010). Research methods for analysing world wide web sustainability communication. *Social and Environmental Accountability Journal*, *30*(1), 26–36.

Luo, Y., Song, K., Ding, X., and Wu, X. (2021). Environmental sustainability of textiles and apparel: A review of evaluation methods. *Environmental Impact Assessment Review*, *86*, 106497.

Luscombe, A., Dick, K., and Walby, K. (2022). Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity*, *56*(3), 1023–1044.

Manshoven, S. Christis, M., Vercalsteren, A.,Arnold, M., Nicolau, M., Lafond, L, Fogh Mortensen, L., Coscieme, L. (2019). Textiles and the environment in a circular economy. European Topic Centre Waste and Materials in a Green Economy.

Martins, A., Branco, M. C., Melo, P. N., & Machado, C. (2022). Sustainability in small and medium-sized enterprises: A systematic literature review and future research agenda. *Sustainability*, 14(11), 6493.

Moldavska, A. and Welo, T. (2019). A holistic approach to corporate sustainability assessment: Incorporating sustainable development goals into sustainable manufacturing performance evaluation. *Journal of Manufacturing Systems*, *50*, 53–68.

Montiel, I. and Delgado-Ceballos, J. (2014). Defining and measuring corporate sustainability: Are we there yet? *Organization & Environment*, *27*(2), 113– 139.

Palma, M., Lourenço, I. C., & Branco, M. C. (2022). Web-based sustainability reporting by family companies: the role of the richest European families. *Accounting Forum, 46* (4).

Pranugrahaning, A., Donovan, J. D., Topple, C., and Masli, E. K. (2021). Corporate sustainability assessments: A systematic literature review and conceptual framework. *Journal of Cleaner Production, 295, 126385*.

Ren, X. (2000). Development of environmental performance indicators for textile process and product. *Journal of cleaner production*, *8*(6), 473–481.

Rovigatti, G. (2017). Production function estimation in R: The Prodest Package. *Journal of Open Source Software*, *2*(18), 371.

Russo, M. V. and Fouts, P. A. (1997). A resource-based perspective on corporate environmental performance and profitability. *Academy of management Journal*, *40*(3).

United Nations, D. o. E. and Development, S. A. S. (2015). Transforming our world: the 2030 agenda for sustainable development.

5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16400

FAIR2: A framework for addressing discrimination bias in social data science

Francisca Garcia-Cobián Richter[1], **Emily Nelson**[1], **Nicole Coury**[1], **Laura Bruckman**[1], **Shanina Knighton**[1,2]
[1]Case Western Reserve University; [2]Center for Infection Prevention and Control Research

Abstract

Building upon the FAIR principles of (meta)data (Findable, Accessible, Interoperable and Reusable) and drawing from research in the social, health, and data sciences, we propose a framework -FAIR2 (Frame, Articulate, Identify, Report) - for identifying and addressing discrimination bias in social data science. We illustrate how FAIR2 enriches data science with experiential knowledge, clarifies assumptions about discrimination with causal graphs and systematically analyzes sources of bias in the data, leading to a more ethical use of data and analytics for the public interest. FAIR2 can be applied in the classroom to prepare a new and diverse generation of data scientists. In this era of big data and advanced analytics, we argue that without an explicit framework to identify and address discrimination bias, data science will not realize its potential of advancing social justice.

Keywords: *Discrimination Bias; Social Data Science Framework; Experiential Knowledge; Causal Diagrams.*

1. Introduction

There is a long, damaging history of purportedly objective use of data that has contributed to reinforcing stereotypes driving fear, isolation, and discrimination in society. Social and health science scholarship of the early 1900s in the United States had an unfortunate role in establishing false connections between crime and the social construct of race (Muhammad, 2019). Prominent social scientists and statisticians used Census Bureau and prison data to make flawed causal claims linking race to crime, helping to cement discrimination in all aspects of society. Mathematician Kelly Miller, sociologist W. E. B. DuBois, and journalist Ida B. Wells were among the Black scholars and activists who counteracted these narratives with carefully crafted arguments hinged on data, logic, and domain-expert knowledge. However, the academic community mostly dismissed this work. Only in 2020 have academic associations issued statements recognizing their lack of understanding of racism and its impact on their work -if not apologizing for the harms caused (American Economic Association, 2020).

In the current era of big data and data technologies, discrimination, reflected in social data in a multiplicity of ways, is a main source of bias. Bias here is generally defined as distortions in inference stemming from data or assumptions. Yet until recently, discrimination in data -particularly discrimination based on the social construct of race- has mostly been ignored by the academic community. Consequently, analyses using these data have little ability to address discrimination or worse yet, may strengthen biased beliefs and perpetuate discrimination. While data analysts may agree that discrimination in society manifests itself in social data, we argue that without an explicit framework to identify and address this problem, data science will not realize its potential of supporting social justice.

In this article, we offer one such framework -FAIR2- that draws from recent literature in the social and health sciences, data science, computer science, community-engaged research practices and metadata principles. It complements the ethical standards of FAIRification (Findable, Accessible, Interoperable and Reusable) principles (Wilkinson et al., 2016) with a set of four additional principles (Frame, Articulate, Identify, Report) specific to working with social data for social impact. Primarily developed to guide students doing data science for social impact, FAIR2 can be more broadly applied to foster stronger communication between researchers, practitioners, and community members whose experiences are represented in the data. Section 2 introduces the FAIR2 framework. Section 3 illustrates the use of FAIR2 with an example of data analytics in the area of public assistance programs. Section 4 provides concluding thoughts.

2. FAIR2

Since the FAIR data principles were proposed in 2016 by a diverse group of stakeholders in academia and industry, their adoption continues to grow. FAIR stands for Findability, Accessibility, Interoperability, and Reusability principles, meant to ease the ability of machines and humans to make informed use of data and maximize the value added of data analytics in a transparent, ethical way (Rocca-Serra et al., 2022). Building on FAIR, we propose an additional set of principles pertinent to social data and analytics, FAIR2. The FAIR2 initials refer to: Frame, Articulate, Identify and Report. This framework recognizes that *data do not speak for themselves*, that observational data reflect discrimination in society, and that an explicit framework to identify and address discrimination biases will further data science's potential to advance social justice.

2.1. Frame

Frame metadata and data with historical context and experiential knowledge of those represented in the data. The inclusion of individuals in administrative data in healthcare, homelessness, policing, among others, is influenced by discrimination. Drawing from the experiential knowledge of people intersecting with these systems will enrich the understanding of who is represented and not represented in the data, how reliable the data are, and what can be learned from it. The Human Rights-Based Approach to Data (OHCHR, 2018) posits that participation by relevant populations in data collection and analysis is key to enhancing the use of data in alignment with international human rights norms. When data has been collected by administrative systems or machines, community participation in establishing the metadata takes on a heightened role. Integrating community knowledge into the metadata can be accomplished via the inclusion of qualitative literature and through designed collaboration meetings -Data Chats- with community members. Data Chats are created with the intention of "center[ing] residents' knowledge, community understanding, and experiences as much as quantitative data (Cohen, Rohan, Pritchard, & Pettit, 2022)". Their structure allows researchers to collaborate and learn from community members while sharing information derived from data. Implementation considerations include developing informed consent forms, planning for accessibility of meeting space and time, sharing a meal, all of which reflects respect and appreciation towards community collaborators. This approach ties in well with the goals of the FAIR2 framework.

2.2. Articulate

Articulate the general model as a causal graph to explicitly state model assumptions (background knowledge) and hypotheses about the role of discrimination in the social phenomenon studied. It has been said that the logic of inference in policy analysis can be

summarized as "assumptions + data → conclusions" (Manski, 2013). In other words, the data do not speak for themselves; assumptions can be a source of subjectivity in inference. Directed Acyclic Graphs (DAGs) are a collection of nodes and directed edges connected under certain conditions to represent a causal model (Pearl, 1995). DAGs make explicit the assumptions embedded in the model, helping to clarify the source of these assumptions (what knowledge and whose knowledge?) and their implications for model estimates (Pearl & Mackenzie, 2018). They can represent the larger context underlying the social phenomenon of interest and the sources of discrimination in outcomes. Furthermore, DAGs facilitate the awareness of selection bias and collider bias that can interfere with the identification of causal effects and interpretation of predictive algorithms. There is much to learn from the recent work of researchers across multiple social science disciplines that have used DAGs to clarify the role of discrimination -in particular racism- on outcomes of well-being (Howe, Bailey, Raifman, & Jackson, 2022), to improve the performance of machine learning (Robinson, Renson, & Naimi, 2020), and to advances in algorithmic fairness (Kilbertus et al., 2017).

2.3. Identify

Identify bias embedded in the data and variables of interest, aiming to minimize bias and report on limitations due to bias. Here we draw from the work of (Kleinberg, Ludwig, Mullainathan, & Sunstein, 2018) and (Lundberg, Johnson, & Stewart, 2021) to systematically analyze potential biases in the estimand and choice of variables used in the model. We set out to answer the following questions: (1) What is the unit-specific quantity of interest -USQ- and the target population? (2) What biases may be introduced by each variable and by using measured versus desired variables? (3) What is the role of variables representing the sensitive attributes (subject to discrimination) in the model? (4) For whom is the USQ missing in the data (selective labels problem; collider bias), why, and how will this affect model estimates?

2.4. Report

Share findings and seek feedback from members or agencies in the community who have experiential knowledge of the social issue analyzed. The increased use of mixed methods research and intersectionality theory (Abrams, Tabaac, Jung, & Else-Quest, 2020) reveal the growing popularity of context-rich and collaborative data. Research has shown that openly communicating and engaging with communities to disseminate data can build trust and ground academic datasets in real world issues (Schalet, Tropp, & Troy, 2020). Sharing data improves health equity, as has been shown in harm reduction strategies for people who use drugs (Salazar, Vincent, Figgatt, Gilbert, & Dasgupta, 2021), and with emergency housing programs (Lane, McClendon, & Matthews, 2017). Engaging community stakeholders in analyzing and reporting findings completes the circuit of utilizing data without losing its human context.

3. Application to an Analysis of Nutrition Assistance Recertification

We illustrate the use of the FAIR2 framework with an analysis of recertification in the Supplemental Nutrition Assistance Program (SNAP), one of the main public assistance programs in the United States. SNAP's predecessor, the Food Stamps Program, was developed in 1933 during the Great Depression to support farmers and people facing food insecurity. Today, SNAP is offered as a food voucher via Electronic Transfer Cards, conditional on meeting low-income thresholds that vary by states. SNAP is among the strongest programs in the US social safety net, with a countercyclical multiplier effect during economic downturns (Canning & Stacy, 2019). States require individuals to follow a recertification process every 6 to 12 months involving detailed proof of income and an interview. Research suggests that up to half of beneficiaries who exit SNAP within their first year were still eligible (Gray, 2019). Failing to recertify despite qualifying -here denoted as FR- is costly to individuals and administrative agencies. *Researchers have sought to study who is affected by FR and what can be done to reduce the rate of FR.*

Frame - Data Chat with SNAP participants who have experienced the recertification process call attention to (1) an under-resourced system: hours long wait times on the phone to make interview appointments, letters requiring additional information that come close to the interview day; (2) a stressful process that can sometimes feel confrontational and the need to "suck up your pride"; (3) a system that seems to penalize any change in employment; (4) within- and across-locality variation in service quality, with higher income localities seemingly performing better than those with higher need and some case workers showing extreme dedication despite the difficulties of navigating an under-resourced system.

Articulate - A common approach to characterize the population experiencing FR is to use administrative data to estimate a regression model for the FR outcome, with demographic and economic characteristics (D&E) as predictors. The left DAG of Figure 1 presents a naive model including "race" and D&E variables compatible with such regression. It implies strong assumptions about the meaning of "race," represented as an individual trait and not directly related to other societal factors that impact FR. The right DAG, explained in the caption of Figure 1, embeds historical and experiential knowledge from our Data Chats and the literature, highlighting systemic issues that are relevant to inform policy. This DAG is inspired by recent work on causal diagrams for studying racial health disparities (Howe et al., 2022; Robinson et al., 2020).

Identify - (1) Let SNAP enrollees subject to recertification in locality L constitute the target population and FR -failure to recertify when eligible- the unit-specific variable of interest. (2) Unable to learn whether enrollees meet eligibility requirements for recertification from administrative data, researchers have sought to flag FR if an individual drops from SNAP and re-enters within 1-3 months (Kenney et al., 2022). This is denoted as churn.

Acknowledging discrepancies between the theoretical and empirical estimand would point researchers to explore re-entry beyond 90 days in the administrative data or through qualitative knowledge. It may also suggest linking administrative data to capture eligibility requirements and thus FR rather than churn episodes. (3) The misconception of "race" as a demographic variable (R in Fig. 1, left DAG) ignores important plausible pathways by which discrimination may be directly impacting FR (HP → CP→ S or FR in Fig. 1, right DAG). While it is valuable to assess differences in FR by racialized groups, acknowledging the underlying mechanisms that lead to inequities in outcomes have strong implications for setting up research designs and identifying policy solutions. (4) Administrative data generated with low resources and characterizing a population facing distress are likely to exhibit non-random missingness issues. Are these patterns consistent with variations in resources or recertification requirements across demographic or geographic groups highlighted in our community conversations?

Report - Co-creating findings allows researchers to engage community members as experts on the social issues they have experienced and share power with the community. Data Chat participants review summaries and conclusions from SNAP recertification analyses and our synthesis of their initial thoughts. They incorporate their input for a final report that will be shared with local administrators of the SNAP program.

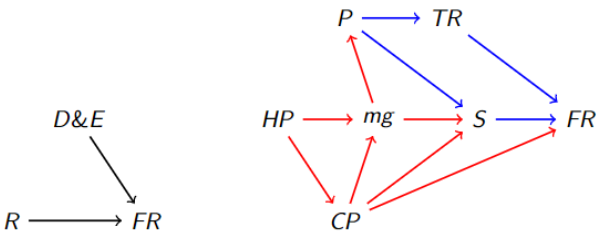


Figure 1. Directed Acyclic Graphs modeling failure to recertify for food assistance (FR). The left graph represents a naive model that aims to explain differences in FR with demographic and economic characteristics (D&E) and race (R). The right model includes historical and experiential knowledge. HP are historical processes that shaped discrimination in US society (like Jim Crow or Redlining); CP are contemporary policies that allocate scarce resources to the social safety net (weak transportation system, understaffed public assistance administration); mg is classification into marginalized grouping; P is poverty; TR is time and resource scarcity; S is stigma.

4. Concluding thoughts

Discrimination in society may influence who is represented and not represented in the data, and how. Variables of interest may be measured with less accuracy in the presence of discrimination, or the metrics used to proxy desired variables may be flawed by

discrimination. Furthermore, modeling assumptions, not always explicitly stated in data analyses, can inadvertently carry biases reflective of discrimination in society. The increased use of scoring algorithms to inform decision making in human services has heightened the need to integrate experiential community knowledge in the development of data technologies for the public interest (Roewer-Despres & Berscheid, 2020). We draw from the work herein cited to build FAIR2 as a tool to address discrimination bias in social data and analytics, strengthen education in data science for social impact, and ultimately propel the field of public interest technology to advance social justice with equity.

Acknowledgements

We thank our community collaborator Ms. Alice Jackson, all Data Chat participants, and the PIT-UN Community-Academic Advisory group at Case Western Reserve University for their insightful contributions. This work was generously funded by grant #015865 from the Public Interest Technology University Network - New America Foundation.

References

Abrams, J. A., Tabaac, A., Jung, S., & Else-Quest, N. M. (2020). Considerations for employing intersectionality in qualitative health research. Social Science & Medicine, 258, 113138. https://doi.org/10.1016/j.socscimed.2020.113138

American Economic Association. (2020, June 5). Statement from the AEA Executive Committee. Retrieved from https://www.aeaweb.org/news/member-announcements-june-5-2020

Canning, P., & Stacy, B. (2019). The Supplemental Nutrition Assistance Program (SNAP) and the Economy: New Estimates of the SNAP Multiplier (Economic Research Report No. 291963). United States Department of Agriculture, Economic Research Service. Retrieved from https://econpapers.repec.org/paper/agsuersrr/291963.htm

Cohen, M., Rohan, A., Pritchard, K., & Pettit, K. L. S. (2022). Guide to Data Chats: Convening Community Conversations about Data. Urban Institute.

Gray, C. (2019). Leaving benefits on the table: Evidence from SNAP. Journal of Public Economics, 179, 104054. https://doi.org/10.1016/j.jpubeco.2019.104054

Howe, C. J., Bailey, Z. D., Raifman, J. R., & Jackson, J. W. (2022). Recommendations for Using Causal Diagrams to Study Racial Health Disparities. American Journal of Epidemiology, 191(12), 1981–1989. https://doi.org/10.1093/aje/kwac140

Kenney, E. L., Soto, M. J., Fubini, M., Carleton, A., Lee, M., & Bleich, S. N. (2022). Simplification of Supplemental Nutrition Assistance Program Recertification Processes and Association with Uninterrupted Access to Benefits Among Participants with Young Children. JAMA, 5(9) https://doi.org/10.1001/jamanetworkopen.2022.30150

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding Discrimination through Causal Reasoning. Advances in Neural Information Processing Systems, 30. Curran Associates, Inc. Retrieved from

https://proceedings.neurips.cc/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. Journal of Legal Analysis, 10, 113–174. https://doi.org/10.1093/jla/laz001

Manski, C. F. (2013). Public Policy in an Uncertain World: Analysis and Decisions. Cambridge, MA: Harvard University Press.

Lane, S. R., McClendon, J., & Matthews, N. (2017). Finding, Serving, and Housing the Homeless: Using Collaborative Research to Prepare Social Work Students for Research and Practice. Journal of Teaching in Social Work, 37(3), 292–306. https://doi.org/10.1080/08841233.2017.1317689

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. American Sociological Review, 86(3), 532–565. https://doi.org/10.1177/00031224211004187

Muhammad, K. G. (2019). The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America, With a New Preface. Cambridge, MA: Harvard University Press.

OHCHR. (2018). A Human Rights-Based Approach to Data. Geneva: Office of the United Nations High Commissioner for Human Rights. Retrieved from Office of the United Nations High Commissioner for Human Rights website: https://www.ohchr.org/sites/default/files/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf

Pearl, J. (1995). Causal diagrams for empirical research. Biometrika, 82(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J., & Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect (1st ed.). USA: Basic Books, Inc.

Robinson, W. R., Renson, A., & Naimi, A. I. (2020). Teaching yourself about structural racism will improve your machine learning. Biostatistics, 21(2), 339–344. https://doi.org/10.1093/biostatistics/kxz040

Rocca-Serra, P., Sansone, S.-A., Gu, W., Welter, D., Abbassi Daloii, T., & Portell-Silva, L. (2022). Reflections on the Ethical values of FAIR. In D2.1 FAIR Cookbook. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.6783564

Roewer-Despres, F., & Berscheid, J. (2020, November 29). Continuous Subject-in-the-Loop Integration: Centering AI on Marginalized Communities. arXiv. Retrieved from http://arxiv.org/abs/2012.01128

Salazar, Z. R., Vincent, L., Figgatt, M. C., Gilbert, M. K., & Dasgupta, N. (2021). Research led by people who use drugs: Centering the expertise of lived experience. Substance Abuse Treatment, Prevention, and Policy, 16(1), 70. https://doi.org/10.1186/s13011-021-00406-6

Schalet, A. T., Tropp, L. R., & Troy, L. M. (2020). Making Research Usable Beyond Academic Circles: A Relational Model of Public Engagement. Analyses of Social Issues and Public Policy, 20(1), 336–356. https://doi.org/10.1111/asap.12204

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Predicting the helpfulness score of videogames of the STEAM platform

Leonardo Espinosa-Leal[1], María Olmedilla[2], Zhen Li[1]
[1]Arcada University of Applied Sciences, Graduate Studies and Research, Finland,
[2]SKEMA Business School – Université Côte d'Azur, France

Abstract

Online reviews comprise a flood of user-generated content, so to identify the most useful reviews is a vital task. As such, many computational models have been made to automatically analyze the helpfulness of online reviews. In this work, we aim to predict the helpfulness score of videogames reviews using an available online dataset of more than 1M rows. We trained three different machine learning algorithms by implementing two strategies, predicting the helpfulness as a regression problem or as a binary classification problem. Our findings show that binary classification is the best method, and the achieved ROC-AUC of the best model is 0.7 with only a selected set of features. In addition, we found that using the feature vectors from a pretrained NLP model does not improve the performance of the models.

Keywords: Videogames; helpfulness; machine learning; NLP; online reviews