CUADERNOS
DE LA FUNDACIÓN

C / 233

Contributions to Risk Analysis: RISK 2022

José María Sarabia
Manuela Alcañiz
Faustino Prieto
Montserrat Guillén
(editors)

Fundación
MAPFRE

Área de Seguro y Previsión Social

Contributions to Risk Analysis: RISK 2022

José María Sarabia
Manuela Alcañiz
Faustino Prieto
Montserrat Guillén
(editors)

Fundación MAPFRE

Selección de ponencias presentadas en el 8º Congreso sobre gestión de riesgos e investigación en seguros-RISK 2022 (Barcelona, 2022), en las que se aborda el análisis de riesgos desde muchas perspectivas diferentes, que van desde el análisis de distribución de la probabilidad hasta el uso de bases de datos masivas. Incluyen modelos clásicos de ciencia actuarial, métodos matemáticos, modelos predictivos o el estudio de las dependencias.

Paseo de Recoletos, 23
28004 Madrid (España)

www.fundacionmapfre.org

ISBN:
Depósito Legal:
Maquetación y producción editorial: Cyan, Proyectos Editoriales, S.A.

PRESENTATION

PRESENTACIÓN

FOREWORD

The 8th Hybrid Workshop on Risk Management and Insurance Research (RISK, 2022) held in Barcelona (Spain) on the 20th and 21st of October, 2022 is an international forum for disseminating recent advances in the field of Risk Analysis, organized by the Research Group on Risk in Insurance and Finance of the University of Barcelona

In line with previous conferences, RISK 2022 provides a platform to share new ideas, research results, and development experiences in actuarial science and finance. In this edition, Prof. Dr. Tony Klein (Queen's Management School, Queen's University Belfast) is the keynote speaker of the inaugural session.

Workshops on Risk Management and Insurance (RISK) have been taking place biannually in Spain since 2005. In 2020, the eighth edition of this workshop should have taken place. However, the COVID-19 pandemic and the public policies necessary to control the crisis made it impossible to hold it. Two years later, things are gradually returning to normal, so we are delighted to celebrate a new edition. More information on the workshop and also on the previous meetings can be found at: https://risk2022.online

There were around 50 submissions. A selection of short papers is included in this book. All of them went through a double peer-review process.

Risk analysis is a field of research that studies the measurement of adverse events, their prevention, and mitigation. The fundamental aspects of the quantitative analysis of risks are fundamentally two: the probability of occurrence of rare phenomena and the severity of the losses. The works published in this volume address this subject from many different perspectives, ranging from the analysis of probability distributions to the use of massive databases. They include classic actuarial science models, mathematical methods, predictive modeling, and the study of dependencies.

The number of research groups in the world working on issues related to risks has not ceased to increase in recent years. Since 2005, a group of Spanish researchers has been meeting once every two or three years to discuss topics related to risk analysis with a high potential for expansion.

The members of the organizing committee would like to thank the scientific committee for their valuable help to make this conference a success. We want to acknowledge participants, presenters, chairpersons, and all authors in general for their contribution to the development of new advances in the analysis of risk. We also want to thank our invited speaker for accepting to play a central role in this conference.

We are indebted to Fundación Mapfre for sponsoring this publication and to all other sponsors for their generous support. In particular, we thank the Universidad de Barcelona for providing us with an excellent location to celebrate the working sessions, we also thank the Santander Financial Institute (SANFI) and the support received from the Spanish Ministry of Economy FEDER grants in projects PID2019-105986GB-C21 and PID2019-105986GB-C22 (Proyectos de I+D correspondientes al Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia) and PANDÈMIES project 2020 PANDE 00074 from AGAUR. Generalitat de Catalunya.

Barcelona, October 2022

EDITORS

José María Sarabia (CUNEF Universidad)
Manuela Alcañiz (Universidad de Barcelona)
Faustino Prieto (Universidad de Cantabria)
Montserrat Guillén (Universidad de Barcelona)

PRÓLOGO

El 8º Congreso sobre gestión de riesgos e investigación en seguros (RISK 2022) celebrado en Barcelona (España) los días 20 y 21 de Octubre de 2022 es un foro internacional para difundir los avances recientes en el campo del análisis de riesgos, organizado por el Grupo de Investigación de Riesgo en Seguros y Finanzas de la Universidad de Barcelona.

En línea con los congresos anteriores, RISK 2022 proporciona una plataforma para compartir nuevas ideas, resultados de investigación y experiencias de desarrollo en ciencias actuariales y financieras. En esta edición, el Prof. Dr. Toni Klein (Queen's Management School, Queen's University Belfast) es el conferenciante principal de la sesión inaugural.

Desde 2005 se celebran congresos de Gestión del Riesgo y Seguros (RISK) bianualmente en España. En 2020 debería haberse celebrado la octava edición de este congreso. No obstante, la pandemia de la COVID-19, así como las políticas públicas que fueron necesarias para controlar la situación, imposibilitaron dicha celebración. Dos años después, las cosas vuelven progresivamente a la normalidad, por lo que estamos encantados de celebrar una nueva edición. Toda la información sobre este evento y las ediciones anteriores, se puede encontrar en: https://risk2022.online

En esta ocasión se presentaron 50 ponencias, de las cuales, una selección se incluye en este libro. Todas ellas han sido seleccionadas a través de un proceso doble de revisión por pares.

El análisis de riesgos es un campo de investigación que estudia la medición de eventos adversos, su prevención y mitigación. Los aspectos fundamentales del análisis cuantitativo de riesgos son fundamentalmente dos: la probabilidad de ocurrencia de fenómenos raros y la gravedad de las pérdidas. Los trabajos publicados en este volumen abordan este tema desde muchas perspectivas diferentes,

que van desde el análisis de distribución de la probabilidad hasta el uso de bases de datos masivas. Incluyen modelos clásicos de ciencia actuarial, métodos matemáticos, modelos predictivos o el estudio de las dependencias.

El número de grupos de investigadores en el mundo que trabajan las cuestiones relacionadas con el riesgo no ha dejado de aumentar en los últimos años. Desde 2005, un grupo de investigadores españoles se ha reunido una vez cada dos o tres años para debatir temas relacionados con el análisis de riesgos con un alto potencial de expansión.

El comité organizador quisiera expresar su agradecimiento al comité científico por su valiosa ayuda para que esta conferencia sea un éxito. Queremos reconocer a los participantes, conferenciantes e investigadores, y todos los autores en general por su contribución al desarrollo de nuevos avances en el análisis del riesgo. Especialmente queremos agradecer a nuestro ponente invitado por aceptar desempeñar un papel principal en este congreso.

Estamos en deuda con Fundación Mapfre por patrocinar esta publicación y con todos los demás patrocinadores por su generoso apoyo. En particular, agradecemos a la Universidad de Barcelona por brindarnos una excelente ubicación para celebrar las sesiones de trabajo, también el apoyo recibido del Ministerio de Economía español. Becas FEDER en los proyectos PID2019-105986GB-C21 y PID2019-105986GB-C22 (Proyectos de I+D correspondientes al Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia) y del proyecto PANDÈMIES 2020 PANDE 00074 de AGAUR. Generalitat de Catalunya.

Barcelona, Octubre 2022

EDITORS

José María Sarabia (CUNEF Universidad)
Manuela Alcañiz (Universidad de Barcelona)
Faustino Prieto (Universidad de Cantabria)
Montserrat Guillén (Universidad de Barcelona)

COMMITTEES

ORGANIZING COMMITTEE CHAIRS

Jose María Sarabia	CUNEF Universidad
Montserrat Guillen	Universitat de Barcelona, RiskCenter-IREA, Spain

LOCAL ORGANIZING COMMITTEE

Manuela Alcañiz	Universitat de Barcelona, RiskCenter-IREA, Spain
Mercedes Ayuso	Universitat de Barcelona, RiskCenter-IREA, Spain
Lluís Bermúdez	Universitat de Barcelona, RiskCenter-IREA, Spain
Helena Chuliá	Universitat de Barcelona, RiskCenter-IREA, Spain
Pablo Fernández-Baldor	Universitat de Barcelona, RiskCenter-IREA, Spain
David Moriña	Universitat de Barcelona, RiskCenter-IREA, Spain
Luiz Ortiz-Garcia	Universitat de Barcelona, RiskCenter-IREA, Spain
Ana M. Pérez Marín	Universitat de Barcelona, RiskCenter-IREA, Spain
Miguel Santolino	Universitat de Barcelona, RiskCenter-IREA, Spain

SCIENTIFIC COMMITTEE

Pilar Abad Romero	Universidad Rey Juan Carlos, Spain
Manuela Alcañiz	Universitat de Barcelona, RiskCenter-IREA, Spain
Jennifer Alonso Garcia	Université Libre de Bruxelles, Belgium
Alejandro Balbás de la Corte	Universidad Carlos III, Spain
Catalina Bolancé	Universitat de Barcelona, RiskCenter-IREA, Spain
Carmen Boado Penas	University of Liverpool, United Kingdom
Enrique Calderín Ojeda	University of Melbourne, Australia
Ana María Debon Aucejo	Universidad Politécnica de Valencia, Spain

ÍNDICE

FRAUD RISK AWARENESS: A NEED OF THE HOUR TO CONTROL RISING FRAUD IN THE ORGANIZATIONS

Ruchi Agarwal
Management Development Institute (MDI), Gurgaon, Haryana, India
ruchi.agarwal@mdi.ac.in

ABSTRACT

Worldwide, organizations are making losing millions of dollars due to rising fraud. Fraud is often understood using casual approaches such as the fraud triangle or the technical aspects of data-driven approaches. Less attention is paid to creating fraud risk awareness. Fraud risk awareness has significant potential to reduce losses as it is a preventive tool to improve the company's risk culture. A field study has been carried out to understand how companies create fraud risk awareness. The research aims to understand the effectiveness of the approach. My research findings revealed that three insurance companies in India developed five fraud risk awareness approaches. The interviews and observations highlighted that two of these approaches did not work well in practice while the other three were successful. The research limitation is the quantitative measurement of awareness level.

1. INTRODUCTION

Globally fraud is rising more than ever. As per industry experts, 10% of claims in the insurance industry are fraudulent. Insurance companies in India are losing almost $6.25 billion annually due to rising fraud. Practically all insurance companies in India are in loss except two. However, there are no published statistics in developing countries like India so far. Rising fraud has become a concern of companies, professional bodies, and regulators.

There are three modes to control fraud. The casual approach is the fraud triangle. The fraud triangle explains the motivation, rationalization and opportunity to commit fraud. However, it does not tell how to control it. Data-driven approaches analyse the large claim data set of insurance companies and highlight red flags. It is like the post-mortem of the data set. A preventive approach to control fraud is missing. There is almost negligible literature on ways to create fraud risk awareness in the organization. In this book chapter, I discussed five approaches to creating fraud risk awareness based on a field study. The first Section will discuss the benefits and challenges, and Section two enumerates approaches to fraud risk awareness and public campaigns which is followed by a conclusion in the last section.

2. BENEFITS AND CHALLENGES OF FRAUD RISK AWARENESS

Awareness of fraud certainly makes a difference. Companies unaware of fraud offer open opportunities for fraudsters to replicate the frauds they are committing in other companies. Fraud risk awareness intends to create speed breakers and hurdles or sometimes elimination of opportunities in the journey of committing fraud. Sometimes, it works as a catalyst in reducing the speed of committing fraud. Aware organizations become the last choice of fraudsters.

In addition, fraud risk awareness has multiple benefits. It helps in prudent decision-making and improves internal controls. The business environment is changing rapidly. Now markets are dynamic, volatile and uncertain. New types of fraud are emerging for which many companies are not prepared. The fraud that occurred within other companies in the industry provides lessons to all companies because it informs why fraud occurred, how it and its consequence. Overall, it supports in streamlining processes and reduction in losses. I have observed many companies closely where companies are known for high growth and recognized as industry leaders, but for years, they are making losses. It's easy to strategically promote fraud by losing internal controls for quick success in the business, while it is rather extremely tough to control later.

Most companies find it challenging to create Fraud Risk Awareness for four reasons. The first and most important reason is a variety of fraud. In developing

countries like India, due to the high population, millions of small-value transactions take place daily while in developed countries, a few high-value transactions take place. Insurance works on the law of large numbers. The high numbers in developing countries make insurance affordable for the masses while increasing fraud makes it loss-making.

Another challenge is the size of fraud they should focus on widespread small, and medium fraud causing increasing losses or single large-scale fraud severely damaging their reputation. An example of small-scale fraud is the insurance company's non-submission of premium cheques and misinforming customers of a lost insurance policy. An example of large-scale fraud may begin with the rationalization of returning money in a short while which never happens. A CEO in charge of one of the biggest coffee chains is overburdened with a non-sharable financial pressure or strain (i.e., loss during gambling). He could not reveal his problems at work or office. After a certain period, the CEO became too stressed; he decided to siphon off 8 million dollars to invest in an all-time booming real estate business to double money in a short while. In the 2008 crisis, he lost money and could not return the money.

The cost of the investigation is another challenging area. Banks and insurance companies engage in many day-to-day transactions. Surprisingly, an insurance company with over 200 employees working in Fraud and Loss Mitigation department can investigate only 0.5% of insurance claims for which the department raised the red alert. Therefore, careful selection of investigation cases is essential. Considering this information, investigating the cost of small, frequent frauds is high and time-consuming. Many times, the claim amount is far less than the cost of the investigation. Information plays a vital role in fraud risk control. For example, fraud in the personal banking business is 2% but fraud in the internet banking business is 20%. In that case, a company may think of prioritizing fraud risk awareness related to the internet banking business.

The final challenge is the non-availability of tools/techniques for fraud risk awareness. Fraud risk information is rarely available in a common pool. There is a lack of expertise. Employees are always interested to hear how fraud is committed but not how it should be controlled. There is high resistance to changing the process, system and policies.

3. APPROACHES TO CREATE FRAUD RISK AWARENESS

I visited three large insurance companies in Asia from 2013 to 2021 to understand fraud risk awareness. The companies initially struggled with the definition of fraud. Fraud significantly varies from fraud risk as one is actuality and the other is a possibility (Power, 2013). Fraud is often confused it with a complaint or mis-selling. In these companies, mis-selling was not considered fraud. I interacted with several executives and observed and took field notes to understand the way fraud risk awareness is created. The results from the field study revealed five bad and good practices in fraud risk awareness.

- **Monthly Magazine:** A large insurance company introduced a monthly Newsletter, which is strictly circulated internally to create fraud risk awareness. To make it interesting, detailed case studies related to fraud are published. Some of the preventive tips were also shared, i.e. how not to fall prey to fraudsters online. Most of the time, fraud risk and mitigation department employees write the cases. However, these employees were well versed with the concepts of fraud risk but faced issues in effectively writing them. In addition, newsletters were edited by employees. The impact of the initiative was unknown to the company whether employees were reading the Newsletter or not.

  After that, this initiative was co-joined with a compulsory quiz on fraud. The company had more than 10,000 employees working at various levels, i.e. agents, managers, executives, leads, vice-presidents and CXOs. The quiz was mandatory for log-in on the PC or laptop, and employees took only 5 minutes to answer the quiz. The initiative was further backed by three-layered certifications, which Chief Risk Officer claimed that he failed in his first attempt to pass certification. Surprisingly, the Human Resource Department designed the quiz, and they set the same difficulty level for an agent and CEO. A condition was added that if every employee does not clear this certification, they will not receive their monthly salary.

  Given this situation, employees decided to cheat the system by hiring a full-time person who could deliver the desired results set by certification and quizzes throughout the year. In this case, all actors were happy. The employees were

receiving salaries on time, the department was able to successfully certify 100% of employees of the company, and they were able to create fraud risk awareness. However, the company is still facing a large number of losses due to the rise in fraud.

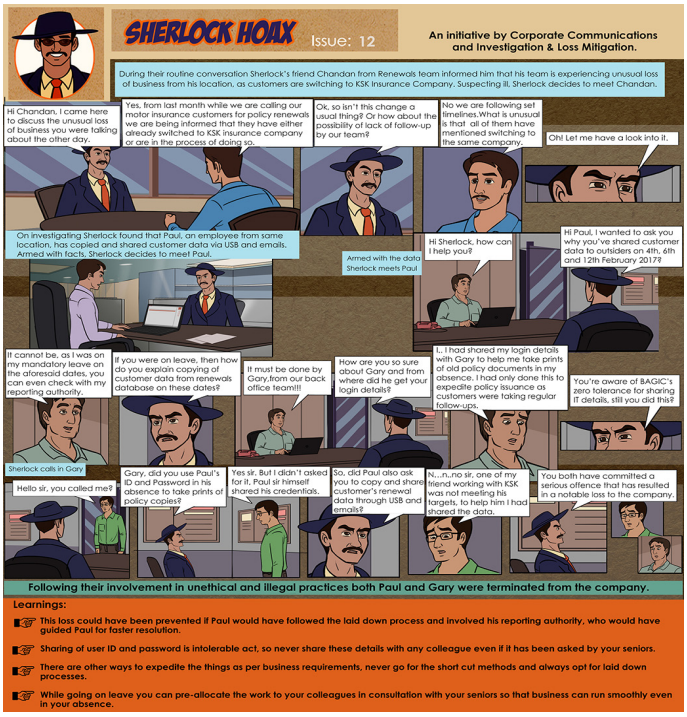
- **Fraud Buster (Institutional Approach):** The company had a large team of over 100 employees in the fraud and loss mitigation department to unearth frauds. The overall objective was to educate its employees to take precautionary measures in case of suspicion of fraud. To achieve this objective, the company's senior officials launched a new initiative, Fraud Buster. Fraud Buster is a one-page internal newsletter with only four sections: background of fraud triggers highlighting the presence of fraud, findings and action taken by the company. The actions taken were used as a tool to prevent fraud in future. The aim was to sensitize staff to various types of fraud risks, and this overall supports loss minimization.

  Fraud Buster was circulated through emails internally. The company also used mechanisms to check whether employees opened the emails. For one year, Fraud Buster was very popular among employees. They were speaking about it in meetings. However, later they lost interest and deleted the emails without reading them.

- **Sherlock Hoax (Creative Approach):** Realizing the lack of interest from employees from Fraud Buster, the same insurance company promoted its senior management and middle management to impart knowledge in the area of Fraud Risk Management in universities and specialized institutes. It had multiple benefits. Curious students participate actively in the workshops, and later during placement, they join the company. The company's executives run a workshop on fraud risk awareness. During the workshop, a student asked, 'Why are fraud risk awareness sessions imparted in a boring and mundane style?' He suggested introducing innovative ways to create fraud risk awareness, such as launching a cartoon series with problems, action and drama. Building upon the suggestion, the company launched a monthly cartoon series that employees loved (see Figure 1). They used to wait for the launch of the new cartoon series. As the audience was adults, due care was taken to set language and content accordingly. The company has been running this series for the last five years. It was one of the successful initiatives in creating fraud risk awareness.

- **Corporate communication (Stakeholder approach):** As some insurance companies were highly active in taking active measures to prevent rising frauds in the organization, they wanted to educate their stakeholders, i.e. customers, employees, vendors, shareholders etc. One of the companies used newspapers to spread the message in the community. Whenever they find a new type of fraud, the company executives deliberately provide information in the print media. This news is also shared internally. This method generated curiosity and showcased the company's strong position in fraud control. The company gained popularity, and it became the last choice for fraudsters. Overall, this created a deterrence effect. This company also developed several tools to control fraud. These tools were discussed in Newsletter to showcase the high level of attention paid by the company to unearth the fraud.

Figure 1. Sherlock Hoax: A Fraud Risk Awareness tool



Source: Shared by the company's official.

- **Forums (Pragmatic approach):** Another Insurance Company was facing one of the highest issues of miss-selling insurance policies. The cause of concern for the company is mis-selling and a spiral of associated issues with mis-selling. For example, when an agent mis-sells an insurance policy, the customer is unhappy and loses trust. Dissatisfied customers often complain to regulators, resulting in penalties, reputation loss, and a long-term legal cost for the company. Agents, in turn, know they mis-sold the policies, and the impact will come in 2-3 years. They try to shift their job, which leads to high employee turnover. The company make losses on several grounds. If a policy lapsed in a short period, overall, it results in huge losses for an insurance company. They lose employees. Employee turnover is another challenge. The customer gets the same policies with higher premiums later. Regulators carry out strict scrutiny and penalise companies for customer complaints. There is a rise in legal cases against companies.

  This insurance company's cause of concern was that the mis-selling was part of the entire industry. If a company tries to stop it, the other company hires the same agents. Later the company found involvement in identity fraud. Multiple policies were taken for the same risks from numerous insurers. The data and findings were not discussed at the industry level, resulting in losses for many. At this stage, fraud risk awareness is needed at the industry level. Over 30 Chief Risk Officers of the Insurance industry formed a community to discuss rising frauds. They found 80 geographical locations and a few specific communities engaged in fraud. Various agents and hospitals involved in the fraud were delisted from all companies altogether. In every meeting, one guest expert is invited to create fraud risk awareness. One brainstorming session is conducted to understand a different point of view to resolve issues. Fraud risk awareness is created within the industry.

4. FRAUD RISK AWARENESS PUBLIC CAMPAIGN

The Association of Certified Fraud Examiners (ACFE) kickstarted International Fraud Awareness Week or Fraud Week campaign in 2000 with an aim to host training opportunities and distribute anti-fraud information and education. Organizations

and individuals registered to officially support prior to International Fraud Awareness Week to host training opportunities. ACFE provides the material to the organizations for the promotion of reasons for rising fraud and ways to control them among employees, stakeholders and local media outlets.

HDFC Bank, one of the biggest private banks in India supported International Fraud Awareness Week, 2021 to increase awareness of all types of fraud and focus on the **importance *of keeping your mouth shut (in Hindi language 'Mooh band rakho'***) to ensure the prevention (See Figure 2). The bank conducted more than 2000 workshops in four months period to safeguard themselves from fraud. Special focus is given to training senior secondary schools and colleges to ingrain awareness in the roots. Some of the messages involved during training are " don't pick up unknown calls; don't click on SMS and mails from strangers. Don't share OTP, card numbers, passwords or PINs etc".

Another non-banking financial service company 'Bajaj Finserv' launched a public awareness campaign in the local language 'Savdhan Rahein, Safe Rahein' (see Figure 3) across digital and social media platforms to educate customers and the public at large. The company wanted to convey an important message to its customer and prospects that strictly avoid paying the refundable advance payment for loans and if customers are buying insurance policies, they should check policy documents carefully. Sometimes fraudsters lure them with low insurance premiums and offer fake policies.

Figure 2. HDFC Bank Fraud Risk Awareness Campaign

Source: Company's website.

Figure 3. Bajaj finserv Fraud Risk Awareness Campaign

Source: Company's website.

5. CONCLUSION

Rising fraud is a crucial issue globally across financial institutions. Fraud risk awareness is essential to create higher cognition of fraud. Five approaches for fraud risk awareness have emerged in a field study in India. My analysis revealed that a few of the approaches failed while others succeeded.

The push and Institutional approach did not work well to improve fraud risk awareness due to a few reasons. First of all, it did not teach employees the importance of the subject. Therefore, they are not engaged. Secondly, employees always think they are doing more. If they don't understand the value of studying fraud risk management, its training is considered an additional burden. Peter Senge, an MIT Associate Professor, said, "the harder you push, the harder the system pushes you back". It is like "the cure can be worse than the disease". When the case company pushed learning to attain attendance for employees, it became a problem. Surprisingly, employees prefer to cheat the system instead of learning fraud risk concepts. Similarly, when reading fraud newsletters has no short-term impact on their earnings or promotion, employees were not willing to invest their time in reading Fraud Buster. The institutional approach of introducing the Newsletter also did not work.

In contrast, employees welcomed creativity in fraud risk awareness. Sherlock Hoax cartoon series introduced by an insurance company became a blockbuster hit, and it motivated employees to understand it like a suspense movie. The company changed their approach to teaching. When they wanted to teach fraud, employees were not interested; however, using creativity, employees were willingly engaged.

The stakeholder approach focused on spreading fraud risk awareness to a larger number of stakeholders using mass media. The company deliberately disclosed what kind of fraud they are able to discover and what were the consequences in print media. These stories created a deterrence effect and reduced the number of widespread frauds. Finally, some firms started participating in the forums where they invited heads of fraud from the industry and experts to learn new tools and techniques to control fraud. The pragmatic approach of

sharing experience, knowledge from experts and learning from other experts improved process gaps.

Overall, fraud risk awareness has been given global recognition to educate stakeholders at multiple levels. Customers need to understand the product (i.e. insurance policies, the procedure for loans) well so that nobody can cheat them easily. Employees need to take preventive measures to reduce or eliminate opportunities to commit fraud. They can bring creativity to make education interesting and more focused to convey the relevant points. Till now, there have been few tools/techniques used to create fraud risk awareness. In future, companies can use the measurement tool to showcase their success to the board, and regulators which would be of interest to a wider audience.

REFERENCES

Agarwal, R. (2018). "A multiple perspective views to rampant fraudulent culture in the Indian insurance industry". International Journal of Indian Culture and Business Management, 16, 4, 416-437.

Agarwal, R., and Kallapur, S. (2021). "Four Ways to Improve Risk Reporting". California Management Review, 63, 52-65.

Amani, F.A., and Fadlalla, A.M. (2017). "Data mining applications in accounting: A review of the literature and organizing framework". International Journal of Accounting Information Systems, 24, 32-58.

Cressey, D.R. (1950). "The criminal violation of financial trust". American Sociological Review, 15, 6, 738-743.

Cressey, D.R. (1953). "Other people's money; a study of the social psychology of embezzlement".

Epstein, B.J., and Ramamoorti, S. (2016). "Today's fraud risk models lack personality". The CPA Journal, 86, 3, 14.

Gray, G.L., and Debreceny, R.S. (2014). "A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits". International Journal of Accounting Information Systems, 15, 4, 357-380.

Law, P. (2011). "Corporate governance and no fraud occurrence in organizations: Hong Kong evidence". Managerial Auditing Journal, 26, 6, 501-518.

Sparrow, M.K. (1998). Fraud Control in the health care industry: Assessing state of the art. US Department of Justice, Office of Justice Programs, National Institute of Justice Washington, DC.

Thornton, D., Brinkhuis, M., Amrit, C., and Aly, R. (2015). "Categorizing and describing the types of fraud in healthcare". Procedia Computer Science, 64, 713-720.

Tomar, D., and Agarwal, S. (2013). "A survey on Data Mining approaches for healthcare". International Journal of Bio-Science and Bio-Technology, 5, 5, 241-266.

LA DEMANDA HOSPITALARIA POR COVID-19: RELACIÓN A CORTO Y LARGO PLAZO CON LA INCIDENCIA REGISTRADA

Manuela Alcañiz
Universitat de Barcelona, RiskCenter-IREA, España
malcaniz@ub.edu

Marc Estévez
Universitat de Barcelona, RiskCenter-IREA, España
marc.estevez.inglada@ub.edu

Miguel Santolino
Universitat de Barcelona, RiskCenter-IREA, España
msantolino@ub.edu

ABSTRACT

La elevada demanda imprevista de atención hospitalaria debida a la pandemia de SARS-CoV-2, sumada a la limitación de los recursos disponibles, ha provocado situaciones cercanas al colapso en los momentos de mayor propagación del virus. Este trabajo modeliza el número de hospitalizaciones a partir del número de casos positivos y del porcentaje de población vacunada en cada momento. A partir de datos españoles del Centro Nacional de Epidemiología, se construye un modelo de corrección del error para introducir la dinámica de corto plazo en esa relación y corregir las desviaciones que se observan a largo plazo. Se obtiene que un incremento del 1% en la población vacunada con pauta completa reduce el riesgo de hospitalización en un 0.8%. El análisis por edades muestra que el virus presenta menor gravedad en los jóvenes, y que la variante Ómicron mejoró las perspectivas de todos los grupos de edad, y especialmente las de los adultos de 50-69 años.

1. INTRODUCCIÓN

La pandemia de SARS-CoV-2 ha supuesto un punto de inflexión para la planificación sanitaria en todo el mundo. La limitación de los recursos hospitalarios, así como el incremento imprevisto de la demanda hospitalaria provocados por el Covid-19 han causado situaciones próximas al colapso en los momentos más críticos. Por este motivo, la literatura ha tratado de proporcionar métodos para predecir la demanda hospitalaria en los distintos momentos de la pandemia, usando a menudo herramientas econométricas.

Algunos autores han utilizado modelos de corrección del error (MCE) para medir el impacto en el sistema sanitario causado por la propagación del SARS-CoV-2 (Nguyen et al., 2021; Mills, 2022). El objetivo de este trabajo consiste en crear un MCE para la relación a largo plazo entre el número de casos positivos por Covid-19 y el número de pacientes hospitalizados debido a la enfermedad en España (Engle y Granger, 1987). La introducción de la dinámica a corto plazo puede corregir las desviaciones del equilibrio que se observan a largo plazo. Además, se desea analizar el impacto sobre el riesgo de hospitalización de la aparición de la variante Ómicron (Tian et al., 2022) y de la evolución de las tasas de población vacunada con la pauta completa. También se realizarán los mismos análisis por grupos de edad. Ello permitirá observar si existen diferencias en la relación entre la incidencia del Covid-19 y las hospitalizaciones en función de la etapa vital del individuo.

El modelo propuesto puede ser útil para mejorar la predicción de la demanda hospitalaria ante situaciones difíciles de anticipar, proporcionando así una valiosa herramienta de planificación sanitaria.

2. DATOS

En este trabajo se han utilizado dos bases de datos distintas. El número diario de casos positivos detectados y los ingresos hospitalarios para cada grupo de edad se han obtenido del Centro Nacional de Epidemiología de España (https://cnecovid.isciii.es). El porcentaje de población vacunada con pauta completa contra el

Covid-19 para los distintos grupos de edad se obtiene de los reportes semanales proporcionados por el Centro Europeo para la Prevención y Control de Enfermedades (https://opendata.ecdc.europa.eu/). Las series cubren el periodo temporal desde el 11 de mayo de 2020 hasta el 20 de marzo de 2022.

Un análisis preliminar muestra la presencia de estacionalidad semanal en las series de casos positivos y de admisiones hospitalarias. Se ha aplicado una transformación logarítmica a ambas series y se ha corregido el efecto de la estacionalidad descomponiéndolas con el método Loess (STL; Cleveland et al., 1990). Por otra parte, la información semanal de la vacunación se transforma en diaria asumiendo que se administran cada día el mismo número de dosis.

3. METODOLOGÍA

Dada la cointegración existente entre las series de casos positivos de Covid-19 y los ingresos hospitalarios, se planteará un modelo de corrección del error (Asteriou y Hall, 2015). Este modelo relaciona el equilibrio a largo plazo entre las dos series temporales con su ajuste a corto plazo, el cual describe cómo reacciona la relación en caso de variaciones en la incidencia de la enfermedad. La relación de equilibrio a largo plazo entre las hospitalizaciones y los casos positivos se representa con la ecuación de cointegración que se muestra a continuación:

$$y_t = b_0 + b_1 x_t + b_2 x_t I_{omic,t} + b_3 z_t + ect \quad (1)$$

donde y_t corresponde al logaritmo de las nuevas admisiones hospitalarias el día t y x_t es el logaritmo de los casos positivos diarios de Covid-19, para $t=1,...,T$, siendo T=679. El efecto de la variante Ómicron en la relación a largo o se analiza con la variable ficticia $I_{omic,t}$, la cual vale 1 en las fechas posteriores al 29 de noviembre de 2021, momento en que se detectó el primer caso de la variante Ómicron en España, y 0 en el periodo anterior. Además, z_t recoge el porcentaje de población vacunada con pauta completa en el tiempo t. Finalmente, el término de corrección del error (ect) se corresponde con los residuos de la regresión.

Dado que los residuos de (1) son estacionarios, se especifica el siguiente modelo de corrección del error para analizar el ajuste a corto plazo y el equilibrio a largo plazo:

$$\Delta y_t = c + \sum_{i=1}^{k} \psi_i \Delta y_{t-i} + \sum_{j=0}^{q} w_j \Delta x_{t-j} + \gamma \cdot ect_{t-1} + \varepsilon_t \qquad (2)$$

En el modelo del corto plazo, las diferencias del logaritmo de las admisiones hospitalarias se explican en función de los retardos de la misma variable dependiente, los retardos de las diferencias del logaritmo de los casos positivos y del primer

retardo del término del error de corrección. El coeficiente γ indica la velocidad de ajuste en el corto plazo cuando ocurre un desequilibrio del largo plazo, es decir, cuando $ect_{t-1} \neq 0$ (Alogoskoufis y Smith, 1991). Finalmente, se realiza un modelo GARCH(1,1) para tratar la heteroscedasticidad de los residuos ε_t, modelizando la varianza como $\sigma^2_t = \alpha_0 + \alpha_1 \varepsilon^2_{t-1} + \alpha_2 \sigma^2_{t-1}$.

4. RESULTADOS

En primer lugar, se analiza el orden de integración de las series de los casos positivos y de los ingresos hospitalarios por Covid-19 (ambas en logaritmo). Como resultado de la prueba de Dickey-Fuller aumentada (ADF; Said y Dickey, 1984) se obtiene que ambas series son integradas de orden 1. Se estima entonces la ecuación de cointegración, cuyos coeficientes se muestran en la Tabla 1.

Tabla 1. Ecuación de cointegración (relación a largo plazo) entre los casos positivos y los ingresos hospitalarios, y prueba de Dickey-Fuller aumentada (ADF).

Coeficiente	Descripción	Valor estimado
b0	Constante	-1.228**
b1	Casos positivos (log)	0.855**
b2	Variante Ómicron	-0.076**
b3	% población pauta completa	-0.008**
ADF	Prueba ADF sobre ect	-5.413**

Nota: **p-valor ← 0.01.

Se detecta que los residuos de la ecuación de cointegración son estacionarios y, por lo tanto, las series de los casos positivos y los ingresos hospitalarios (ambas en logaritmo) están cointegradas. En referencia a los coeficientes estimados de la relación a largo plazo, se observa que un aumento del 1% en los casos positivos detectados supone un aumento del 0.855% de los ingresos hospitalarios, disminuyendo este valor a partir de la irrupción de la variante Ómicron al tener su coeficiente un signo negativo. Además, se obtiene que un aumento en el porcentaje de población vacunada con pauta completa supone un descenso en los ingresos hospitalarios.

Una vez estimada la relación a largo plazo, se define el modelo de corrección del error (2) con la heteroscedasticidad corregida mediante un modelo GARCH(1,1). La selección del orden *(k,q)* del modelo se realiza a partir del criterio de información bayesiano (BIC; Schwarz, 1978). Como resultado, se obtiene un modelo con 11 retardos para las diferencias del logaritmo de los ingresos hospitalarios y un retardo para las diferencias del logaritmo de los casos positivos. Los resultados se muestran en la Tabla 2.

Se puede observar que el coeficiente relacionado con el término de la corrección del error es significativo y negativo, condición necesaria para que la relación entre los casos positivos y los ingresos hospitalarios tienda hacia el equilibrio. El valor del coeficiente indica que cerca de un 9% de cualquier desequilibrio ocasionado es corregido al cabo de un periodo.

Tabla 2. Modelo de corrección del error (relación a corto plazo) entre las series de casos positivos y de ingresos hospitalarios (en escala logarítmica)

Coeficiente	Descripción	Valor estimado
c	Constante	-0.003
ψ_1	Retardo 1 diferencias hospitalizaciones	-0.591**
ψ_2	Retardo 2 diferencias hospitalizaciones	-0.298**
ψ_3	Retardo 3 diferencias hospitalizaciones	-0.074
ψ_4	Retardo 4 diferencias hospitalizaciones	0.009
ψ_5	Retardo 5 diferencias hospitalizaciones	0.115*
ψ_6	Retardo 6 diferencias hospitalizaciones	0.160**

Coeficiente	Descripción	Valor estimado
ψ_7	Retardo 7 diferencias hospitalizaciones	0.349**
ψ_8	Retardo 8 diferencias hospitalizaciones	0.232**
ψ_9	Retardo 9 diferencias hospitalizaciones	0.183**
ψ_{10}	Retardo 10 diferencias hospitalizaciones	0.179**
ψ_{11}	Retardo 11 diferencias hospitalizaciones	0.105**
w_1	Diferencias casos positivos	0.304**
w_2	Retardo 1 diferencias casos positivos	-0.047**
Corrección del error		
γ	Término de corrección del error	-0.088**
Ecuación de la varianza		
α_0	Constante	$1.6 \cdot 10^{-4}$**
α_1	Término de error	0.110**
α_2	Término de varianza	0.866**
AIC	AIC del MCE	-2.169
BIC	BIC del MCE	-2.049
HQ	HQ del MCE	-2.123
R^2	Coeficiente de determinación del MCE	0.425

Nota: **p-valor ← 0.01; p-valor ← 0.05; AIC = criterio de información de Akaike; BIC = criterio de información bayesiano; HQ = criterio de información de Hannan-Quinn.

Tabla 3. Modelo de corrección del error entre los casos positivos y los ingresos hospitalarios (en escala logarítmica) por grupos de edad

Coef.	Descripción	Valores estimados			
		20-49	50-69	70-79	80+
c	Constante	-0.001	-	-	0.001
ψ_1	Retardo 1 diferencias hospitalizaciones	-0.665**	-0.637**	-0.776**	-0.548**
ψ_2	Retardo 2 diferencias hospitalizaciones	-0.316**	-0.266**	-0.433**	-0.234**
ψ_3	Retardo 3 diferencias hospitalizaciones	-	-	-0.162**	-
w1	Diferencias casos positivos	0.301**	0.365**	0.410**	0.314**
w2	Retardo 1 diferencias casos positivos	-	-	-	0.091**
Corrección del error					
γ	Término de corrección del error	-0.076**	-0.072**	-0.112**	-0.112**

Coef.	Descripción	Valores estimados			
		20-49	50-69	70-79	80+
Coeficientes largo plazo					
c0	Constante	-1.970**	-2.198**	-1.372**	-1.004**
b1	Casos positivos (log)	0.829**	1.005**	1.005**	0.998**
b2	Variante Ómicron	-0.114**	-0.146**	-0.133**	-0.117**
b3	% población pauta completa	-0.005**	-0.005**	-0.004**	-0.001*
Ecuación de la varianza					
α_0	Constante	4·10-4*	5·10-4**	4·10-4*	2·10-4*
α_1	Término de error	0.126**	0.110**	0.126**	0.121**
α_2	Término de varianza	0.861**	0.866**	0.868**	0.878**
AIC	AIC del ECM	-0.858	-1.126	-0.717	-1.008
BIC	BIC del ECM	-0.804	-1.072	-0.656	-0.948
HQ	HQ del ECM	-0.837	-1.105	-0.693	-0.985
R2	Coeficiente de determinación del ECM	0.436	0.409	0.478	0.29

Nota: **p-valor ← 0.01; *p-valor ← 0.05; AIC = criterio de información de Akaike; BIC = criterio de información bayesiano; HQ = criterio de información de Hannan-Quinn.

A continuación, se ajusta un modelo de corrección del error para cada uno de los siguientes grupos de edad: 20-49, 50-69, 70-79, y 80 años y más. Se desea observar si existen diferencias en la magnitud del efecto de los casos positivos y de la vacunación sobre el número de ingresos hospitalarios en función de la edad. Nuevamente, se selecciona el número de retardos a utilizar en cada modelo basándose en el BIC. Los resultados obtenidos se muestran en la Tabla 3.

En referencia a los coeficientes de la ecuación de largo plazo, se constata que en todos los grupos un incremento en los casos positivos supone un aumento en los ingresos hospitalarios, siendo este menor para el grupo más joven. Además, en todos los segmentos de edad se reduce el número de positivos en Covid-19 que requieren hospitalización a partir de la variante Ómicron, especialmente para el grupo de 50-69 años. A medida que avanza el porcentaje de población con pauta completa, el número de hospitalizaciones disminuye para todas las edades, siendo este efecto menos intenso a partir de los 80 años.

5. CONCLUSIONES

En este trabajo se analiza y cuantifica la relación entre el número detectado de casos de Covid-19 y el número de ingresos hospitalarios por esta enfermedad en España entre mayo de 2020 y marzo de 2022. Las relaciones a largo y a corto plazo se descomponen mediante un modelo de corrección del error, cuya heteroscedasticidad es corregida mediante un proceso GARCH. Se observa que la relación entre las dos series temporales es estable y tiende hacia un equilibrio a largo plazo. Se muestra que, tras un desequilibrio en la relación a largo plazo, alrededor de un 9% del desequilibrio es corregido tras un periodo.

Se analiza también el impacto de dos variables distintas que interactúan en la relación a largo plazo entre los casos positivos y los ingresos hospitalarios: la tasa de vacunación con pauta completa y la aparición de la variante Ómicron. Se obtiene que un incremento del 1% en la población vacunada con la pauta completa reduce el riesgo de hospitalización en un 0.8%. Además, en línea con Ulloa et al. (2022), se obtiene que a partir de la aparición de la variante Ómicron un aumento de la incidencia supone un menor incremento en el número de hospitalizaciones con respecto a las anteriores variantes.

Con el fin de observar posibles diferencias en la relación entre casos positivos e ingresos hospitalarios en función de la edad, se ha analizado el riesgo de hospitalización para distintos grupos de edades. Como resultado, se obtiene que un aumento en los casos positivos afecta en menor medida al riesgo de hospitalización de los más jóvenes, como también constatan Palmer et al. (2021). Además, la variante Ómicron ha supuesto una disminución de la gravedad de la infección en todos los grupos de edad, y especialmente en el de 50-69 años. En consonancia con los resultados de Müller et al. (2021), se concluye que la vacunación también ha ayudado a reducir las hospitalizaciones en todos los segmentos de edad, siendo menos efectiva a partir de los 80 años.

AGRADECIMIENTOS

Este estudio se ha financiado con la ayuda recibida de la Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya mediante el proyecto 2020-PANDE-00074. También agradecemos el apoyo del Ministerio de Ciencia, Innovación y Universidades a través del proyecto PID2019-105986GB-C21.

REFERENCIAS

Alogoskoufis, G., and Smith, R. (1991). "On error correction models: specification, interpretation, estimation". Journal of Economic Surveys, 5, 1, 97-128.

Asteriou, D., and Hall, S.G. (2015). "Applied Econometrics". 3rd ed. Basingstoke: Palgrave Macmillan.

Cleveland, R., Cleveland, W., McRae, J., and Terpenning, I. (1990). "STL: A seasonal-trend decomposition procedure based on Loess". Journal of Official Statistics, 6, 3-73.

Engle, R.F., and Granger, C.W.J. (1987). "Co-integration and error correction: representation, estimation, and testing". Econometrica, 55, 251-276.

Mills, T.C. (2022). "Modelling the link between Covid-19 cases, hospital admissions and deaths in England". National Accounting Review, 4, 1, 38-55.

Müller, L., Andrée, M., Moskorz, W., Drexler, I., Walotka, L., Grothmann, R., et al. (2021). "Age-dependent Immune Response to the Biontech/Pfizer BNT162b2 Coronavirus Disease 2019 Vaccination". Clinical Infectious Diseases, 73, 11, 2065-2072.

Nguyen, H., Turk, P., and McWilliams, A. (2021). "Forecasting COVID-19 hospital census: A multivariate time-series model based on local infection incidence". JMIR Public Health and Surveillance, 7, 8, e28195.

Palmer, S., Cunniffe, N., and Donnelly, R. (2021). "COVID-19 hospitalization rates rise exponentially with age, inversely proportional to thymic T-cell production". Journal of The Royal Society Interface, 18, 176, 20200982.

Said, S.E., and Dickey, D.A. (1984). "Testing for unit roots in autoregressive-moving average models of unknown order". Biometrika, 71(3), 599-607.

Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics, 6, 2, 461-464.

Tian, D., Sun, Y., Xu, H., and Ye, Q. (2022). "The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant". Journal of Medical Virology, 94, 6, 2376-2383.

Ulloa, A.C., Buchan, S.A., Daneman, N., and Brown, K.A. (2022). "Estimates of SARS-CoV-2 Omicron variant severity in Ontario, Canada". JAMA, 327, 13, 1286-1288.

PREDICTIVE ANALYTICS FOR PAID-UPS IN LIFE INSURANCE SAVING PRODUCTS

David Anaya Luque
Universitat de Barcelona, PhD student at the Faculty of Economics and Business, España
danayalu7@alumnes.ub.edu

Lluís Bermúdez
Universitat de Barcelona, Dept. Matemática Económica, Financiera y Actuarial, Riskcenter-IREA, España
lbermudez@ub.edu

Jaume Belles-Sampera
Riskcenter-IREA, Grupo Catalana Occidente, España
belles.sampera@ub.edu

ABSTRACT

Life insurance companies are subject to several risks related to savings products. One of these risks emerges when policyholders exercise, at some point in time before maturity, the option of stopping the payment of the regular premiums initially agreed for the whole life of the policy. This risk is commonly known as paid-up risk. This paper contributes to the analysis of this particular risk, by means of comparing five predictive models to guess the future state of policies in force regarding premium payment: paid-up or active. The models analysed belongs to the following families: logistic regressions, classification decision trees, Random Forests, XGBoost and neural networks. The results show that Random Forest and XGBoost models outperformed the other classification algorithms analysed in this study. As a relevant contribution, each model is also evaluated from a decision-making perspective, in which the ability of each model for identifying client profiles that lead to reduce the overall paid-up probability is checked.

1. INTRODUCTION

A universal life product is an insurance product offering the policyholder a wealth account, where paid premiums as well as returns on the account value are accumulated therein. If it is a savings product, then the return is based on a contractual interest rate guaranteed by the insurer, which can be complemented with some profit sharing. If it is a unit-linked product, then the return is based on the performance of the underlying investment fund selected by the policyholder, and typically there is no profit sharing option.

Apart from the wealth account, the policyholder is provided with some protection guarantees, typically for death or disability events. These additional guarantees allow the beneficiaries to get an extra amount of money over the value of the wealth account if any of those adverse events occur. The policyholder has several options, which can be exercised at any time along the policy life-time: can stop paying regular premiums (or increase or decrease the target amount of the regular premium); can pay supplementary premiums, apart form the regular ones; can make partial or total withdrawals; can transfer the wealth account to other investment products.

From these optionality, two examples of risks related to policyholder behaviour can be derived: lapse risk and paid-up risk. In the case at hand, the paid-up risk emerges when policyholders exercise, at some point in time before maturity, the option of stopping the payment of the regular premiums initially agreed for the whole life of the policy.

In terms of negative impacts for the insurer, lapse risk is certainly more dangerous than paid-up risk. This is probably the reason why there is much more literature on lapse risk and the methodologies to predict and mitigate it than the literature on payment risk. For example, from an insurer's perspective, the excess policy with lapse status

The next sections introduce the random variables to be studied, present the set of predictive models developed and list the valuation metrics to compare the performance of the models. At the same time, a description of the dataset analysed is

given, while the results obtained are presented and discussed in the results and conclusions section. A relevant part of this paper focuses on taking the analysis and discussion of each model a step further: it raises questions about the ability of each model to provide segmentations of customers or customer profiles most likely to exercise the paid options. This last section is developed further in the full paper.

2. RARE EVENTS

Balanced data refers to the concept where both (or multi) classes contain about equally many observations. In both binary classification problems and multi-class classification problems, the sensitivity of the distribution of the data may alter the results, as imbalanced data can lead to classifiers, which are biased towards predicting the majority class, to provide inaccurate and incomplete information about the quality of the model, (Maaloouf and Siddiqi, 2014; Wallace and Trikalinos, 2011).

To address the problem of biased classifiers on imbalanced data and mitigate the possible negative impact of rare events, two main techniques are currently proposed in the literature: cost-sensitive learning and the resampling techniques, (Elkan, 2001; Ling and Sheng, 2008, Thai-Nghe and Schmidt-Thieme, 2010; Khan and Togneri, 2017; Estabrooks, 2004).

Therefore, in addition to the initial predictive models to be considered during the study, an attempt will also be made to address the problem of unbalanced data and the possible effects on the derivation of predictive calculations.

We will focus especially on analysing the behaviour of logistic regression for classification problems, adding the logit function to prefix the output between 0 and 1. We will also focus on algorithms that use the loss function of our base model, such as decision trees, or as within the same family of ensemble models, and with similar functionalities, algorithms such as Random Forest and XGBoost because of the good performance they show with classification problems, (Lin et al., 2021; Dhieb, et al., 2019). Finally, because of the similarity in terms of the search for the minimum of the loss function, but from an individual prediction perspective and

not with the predictions of multiple models (boosted), neural networks are also introduced.

Under these premises, this study will be focused on analysing the consequences of the use of each of the techniques specified above for the treatment of imbalanced data. In addition, it is compared, with respect to the original data, whether there is any significant impact on the results obtained from each of the models using the AUC-ROC curve metric as the main indicator, which model reports better metrics and whether the application of data balancing methods improves the performance of the results obtained.

3. DATA

The data contains 593,464 life insurance policies sold during the period from 2011 to 2020, specifically permanent life insurance products (Universal Life products) with a detailed description of each policyholder's characteristics, premium payment status and the volume of new production portfolio for each year included in the history.

Even though the database contemplates a 10-year history, we are going to focus on selecting only the years 2018 and 2019 (portfolio of policyholders in each calendar year period), with the objective of predicting the one-year premium payment probability.

Therefore, the resulting database will contain the same policy in both periods, but those time increment variables, such as capital, premium or fund value, will be different. In this case, we start from the initial state of 2018 and select the variables that are susceptible to temporary increase from one period to the other in 2019.

For the numerical variables, we have the variable *cap*, which is the capital sum in the event of death. The variable *res*, which is the current value of the funds. The variable *prem*, which is the initial value of the premium paid. The variable *loy*, which is the period between the inception date and the calculation date. The variable *age*, which is the current age of the insured in years and, finally, the variable

finalpp, which is the remaining number of years of premium payment according to the policy contract.

For the categorical variables, we have the variable *incr*, which is the rate of premium increase. The variable *aggprod*, which is the grouping of the different savings products available in the portfolio. The variable *sex*, which is the gender of the insured. The variable *freq*, which is the frequency of premium payment and, finally, the variable *suspb*, which is whether the policy has previously suspended premium payment in the past.

The following tables present the descriptive statistics for the explanatory variables available for the study.

Table 1. Description of continuous variables

Variable	Mean	Min.	Pctl.25	Pctl.50	Pctl.75	Max.
cap	4,619.10	0.00	887.90	5,327.40	5,336.40	155,382.50
res	9,232.00	0.00	1,104.00	3,331.00	8,861.00	2,523,191.28
prem	912.70	0.00	532.74	639.29	1,065.48	31,964.45
loy	5.24	0.00	1.00	4.00	7.00	35.00
age	47.43	14.00	40.00	47.00	55.00	90.00
finalpp	15.80	0.00	6.00	11.00	20.00	83.00

4. RESULTS

Once the database with the two selected years has been filtered, we code all the categorical variables previously mentioned. A minimum-maximum scaling is applied to scale all policies in the range [-1,1]. In parallel, the rest of the numerical variables have been preprocessed before being incorporated into the runs of each model.

Additionally, to ensure the robustness of the models run and the accuracy of the results, 5 repetitions of 10-fold increased validation are established - 10-fold increased validation of the training data 5 times, using a different set of folds for each increased validation in order to obtain more accuracy and robustness of the increased validation tests.

The final result is reported as the average precision of the cross-validation, both for the main metrics of interest and, for example, for the construction of the confusion matrix itself.

Table 2. Description of categorical variables

Variable	Categories
incr	Geom (38,354-32,22%); Const (2,917-2,45%); Arithm (77,780-65,33%)
aggprod	AG1 (30,190-25,36%); AG2 (23,766-19,96%); AG3 (49,298-41,41%); AG4 (15,797-13,27)
sex	Male (53,374-44,83%); Female (65,677-55,17%)
freq	Yearly (12,369-10,39%); Monthly(97,685-82,05%); Other (8,997-7,56%)
suspb	Yes (12,356-10,38%); No (106,695-89,62%)

From the cross-analysis performed for each of the models, according to the type of transformation applied during the process of obtaining the predictions, a series of results are derived in order to be able to compare according to metrics of interest which model and which data adjustment shows optimal results. Additionally, these results are also compared without applying any transformation versus the transformed data according to the cost-sensitivity learning or resampling analysis methods. As a result, it can be observed that the specificity when the data is not transformed tends to fall, and that is because the data, being unbalanced, the cut-off level to determine the classification of each of the predictions must be rescaled according to the current volume of 0 and 1 in the database.

Table 3. Results of the three best models (undersampling)

	Logistic regression		Random Forest		XGBoost	
	Act. (0)	Act. (1)	Act. (0)	Act. (1)	Act. (0)	Act. (1)
Pred. (0)	7,493	1	7,738	6	7,798	21
Pred. (1)	1,219	332	974	327	914	312
NPV	99.99%		99.92%		99.73%	
PPV	21.40%		25.10%		25.43%	
Youden Ind.	85.68%		86.88%		83.17%	
F1-Score	35.24%		39.97%		40.00%	
Accuracy	86.51%		89.16%		89.66%	
B. Accuracy	92.84%		93.44%		91.58%	

Therefore, when applying weights (penalties) or resampling techniques, automatically the adjustment of the volume of 0 and 1 in the portfolio is around 50% and, therefore, the results are rebalanced on this level.

In our analysis, it is determined that the best technique that shows the best results is to undersample the data. The following table lists the three best models and the main metrics considered as selection criteria beyond AUC-ROC, sensitivity and specificity.

5. CONCLUSIONS

This paper has presented an additional insight beyond the traditional methods of rare event detection and classification especially in the field of insurance. In the same way, it has been observed how, beyond the impact that can be derived from unbalanced data, the non-transformation of the data before carrying out the study with the different predictive models selected, could be considered erroneously results that do not fit reality.

Therefore, two main conclusions can be drawn. The first is the need for adjustment and transformation of the data when working with two where the volume of occurrence and non-occurrence events differ significantly. In this study, the volume of positive events in the database was 4%, while negative events were 96%, which caused the results to skew towards zero.

By applying techniques such as cost-sensitivity analysis and resampling techniques, we obtained the best fit for our data, which is undersampling.

From there, through the application of comparative metrics linked mainly to the ROC curve, the area under the ROC curve and the sensitivity and specificity, each of the predictive models analysed is compared. It is obtained that for ease of application, the logistic regression model performs very well, while the Random Forest model and the optimised approximation of the gradient boosted decision tree algorithm (XGBoost), are the ones that perform best in terms of results.

As a conclusion, it is possible to derive the probability of premium payment by analysing individually the characteristics of each policyholder. Furthermore, it opens the door to the fact that this study is not only reserved for this hypothesis (within the persistence hypotheses) but can be extrapolated to the rest of the hypotheses considered in the valuation of life insurance products. Also, it should be noted that it is not only limited to savings products in the field of insurance, but can be extended to other financial and banking fields, where previous studies already exist.

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Science, Innovation and Universities (grant PID2019-105986GB-C21). Likewise, we want to acknowledge the support received by AGAUR of the Catalan Government (Grants 2020DI9 and 2017SGR1147).

REFERENCES

Bauer, D., Gao, J. Moenig, T. Ulm, E.R., and Zhu, N. (2015). "Policyholder exercise behavior in life insurance: The State of Affairs". In: Risk Management and Insurance Faculty Publications 1. url: https://scholarworks.gsu.edu/rmi_facpub/1.

Biagini F., Huber T. Jaspersen, J.G., and Mazzon, A. (2019). "Estimating extreme cancellation rates in life insurance". In: Munich Risk and Insurance Center Working Paper 33.

Bradley, A.P. (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". Pattern Recognition, 30, 7, 1145-1159.

Changki, K. (2005). "Modeling surrender and lapse rates with economic variables". North American Actuarial Journal, 9, 4, 56-70.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). "SMOTE: synthetic minority over-sampling technique". Journal of Artificial Intelligent Research, 16, 321-357.

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system". In Proceedings of the 22nd ACM Sigkdd international conference on knowledge discovery and Data Mining, 785-794.

Elkan, C. (2001). "The Foundations of Cost-Sensitive Learning". In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, 973-978.

Frees, E., Derrig, R., and Meyers, G. (Eds.) (2014). Predictive modeling applications in Actuarial Science. Cambridge University Press International Series on Actuarial Science.

Hand, D. (2012). "Assessing the performance of classification methods". International Statistical Review, 80, 400-414.

He, H., and Garcia, E.A. (2009). "Learning from imbalanced data". IEEE Transactions on knowledge and data engineering, 9, 21, 1263-1284.

López, V., Fernández, A., García, S., Palade, V., and Herrera F. (2013). "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics". Information Sciences, 250, 113-141.

TWO MULTI-POPULATION MORTALITY MODELS: A COMPARISON OF THE FORECASTING ACCURACY WITH RESAMPLING METHODS

Ana Debón
Universitat Politècnica de València, España
andeau@eio.upv.es

David Atance
Universidad de Alcalá, Departamento de Economía y Dirección de Empresas, España
david.atance@uah.es

ABSTRACT

Forecasting mortality usually serves practical needs because its improvements could have enormous social and financial implications. The main objective of this paper is to compare multi-population mortality models for projecting mortality rates in regional dynamic life tables. Our contribution compares different approaches for adjusting regional mortality rates using resampling methods. Using official data from the Spanish National Institute of Statistics (Instituto Nacional de Estadistica, INE), we applied modifications of the Lee-Carter model to Spanish regions. Consequently, we propose a procedure that can be applied to the related life tables database, allowing us to choose the most appropriate model for any geographical area.

1. INTRODUCCIÓN

The improvements in longevity benefit individuals and society. However, the effect of demographic changes provides social and financial risk in critical areas that imply an answer for the governments and companies. To solve this issue, an

accurate valuation of future mortality could help the institutions and companies take the appropriate actions.

Lee and Carter (1992) construct one of the most well-known and applied methods in the demographic and actuarial fields. This method is a two-factor model with fixed age and period effects developed to fit and forecast one population. Since its publication, the model has inspired many variants and extensions to improve the fitting and forecasting accuracy and jointly model mortality of several related populations.

In the context of modeling two or more related populations, the literature focuses on mortality projections for specific areas/regions of a global population. In this regard, two variants of the original Lee-Carter fit mortality in regions that form part of a group rather than considering them individually. On one side, Russolillo et al. (2011) propose adding a new multiplicative effect to represent different countries/regions. On the other side, Debón et al. (2011) add the region/country effect as an additive index.

The accuracy of the fitting and prediction of mortality from a group belonging to a larger population is a fundamental tool to provide an adequate solution for studying the evolution and the mortality of a group with a small sample. Nowadays, the explosion of "big data" has brought many problems, and machine learning has appeared as a new subfield in statistics, focused on supervised and unsupervised modeling and prediction. Within machine learning, "resampling methods" are a fundamental tool to test the model's performance out-of-sample (Gareth et al., 2013). Thus, these methods can be employed to evaluate the forecasting ability of two multi-population mortality models. Therefore, the objective of this paper is to use machine learning techniques to evaluate the predictions of those two multi-population mortality models.

2. MULTIPOPULATION MORTALITY MODELS

This section deals with multi-population mortality models for different territories. The main objective is to forecast the actual mortality rates $q_{x,t,i}$ for each age x, period t, and region i.

2.1. Multiplicative mortality model

Russolillo et al. (2011) propose adding a multiplicative index term to shift the mortality for each region in the group of populations. That is:

$$\operatorname{logit}\left(q_{x,t,i}\right)=\log\left(\frac{q_{x,t,i}}{1-q_{x,t,i}}\right)=\alpha_x+b_x k_t I_i+\varepsilon_{x,t,i}; \qquad (1)$$

where α_x, b_x and are the common age-dependent and time-dependent parameters for all the considered regions. k_t is an index that describes the general trend of the group of populations over time. Meanwhile, a multiplicative index I_i represents the differences in mortality among regions.

2.2. Additive mortality model

Debón et al. (2011) incorporate an additive index term to modify the mortality of each region in the multi-population model. Its expression is:

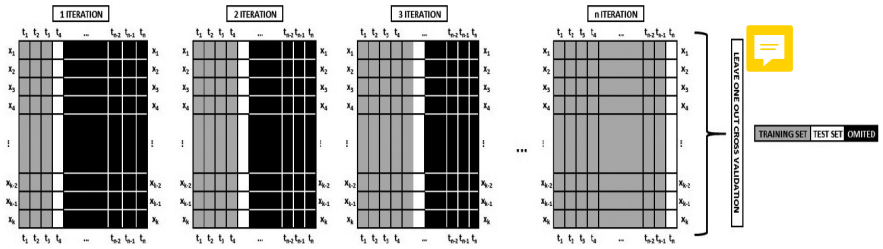
$$\operatorname{logit}\left(q_{x,t,i}\right)=\log\left(\frac{q_{x,t,i}}{1-q_{x,t,i}}\right)=\alpha_x+b_x k_t+I_i+\varepsilon_{x,t,i}; \qquad (2)$$

The number of the parameters is the same as in the model of Russolillo et al. (2011), and the parameters have similar interpretations. Debon et al. (2001) can be consulted details of comparison, but it has a more straightforward structure than the previous one because it adds the region as an additive index.

3. RESAMPLING METHOD.

Resampling methods are used to evaluate the model forecasting accuracy. These methods repeatedly draw samples from the dataset, train (fit), and test the model on each sample (Atance et al., 2020). When the model is run multiple times, additional information is obtained that cannot be acquired by running the model only once.

Figure 1. Schematic display of the resampling methods for time series. The training, validation, and omitted sets are gray, white, and black, respectively.



There are several types of resampling methods in the literature, although, in this paper, we only apply leave-one-out cross-validation (LOOCV). According to this method, only one year onwards is predicted starting from several years' data, incorporating that year in the next iteration (Figure 1). This procedure is repeated n times (n being the number of years), and the measure of the quality of the forecast is obtained as,

$$\text{Leave-One-Out Cross-Validation}_n = \frac{1}{n}\sum_{i=1}^{n}\text{Goodness of fit measures}_i; \qquad (3)$$

4. APPLICATION TO MORTALITY DATA

The data employed in this study were downloaded from the Spanish National Institute of Statistics (INE). INE has provided life tables for Spain and all the regions since 1991. We have included 17 of the 19 Spanish regions in the analysis. Two of them, Ceuta and Melilla, were eliminated due to their tiny population size. It should be mentioned that for Spain, INE supplies complete life tables; meanwhile, regions are abridged with the 5-years age group. Thus, we use abridged life tables to ensure the uniformity of all mortality data. The sample period covers 1991 to 2020, and the age range from 0, 1-4, 5-9, ... to 90-94 (The group 95-99 is also available in the life tables. Although we do not include for maintain the quality of the data). The life tables were downloaded from INE in PC-Axis file format. We used pxR R-package to read this file because the fitting was conducted by R Core Team (2022).

The steps of the LOOCV method are:

1. Set the period 1991-2000 for training the model.

2. Thus, the multi-population mortality models in all the regions are fitted in that period, applying equations (1) and (2).

3. Projection of kt value one year ahead using ARIMA models. Then, the probabilities of death for all ages in each region studied are obtained for that year.

4. The procedure is repeated 20 times, and in each iteration, the training set incorporates one additional year, and the test set moves one year ahead.

5. Afterward, we estimate the goodness-of-fit measure to test the forecasting accuracy of the model.

6. This paper uses the root of the mean squared errors (RMSE) as the global measure for forecasting accuracy. Table 1 shows the results for all ages in every studied region. Also, Table 1 includes the outcomes for all regions as "total". Boldface numbers are the model with the best out-of-sample measure of accuracy.

Based on the results, there are no huge differences between both models. Nevertheless, the additive model works better in more regions than the multiplicative mortality model (Table1). Also, it is interesting to observe that the RMSE with englobes in all regions of Spain has a higher forecasting accuracy in females than males. An outcome is shown by Dong et al. (2020) and is confirmed in this work. This fact shows that the region index improves men's adjustment more than women's.

Furthermore, to show how each model works along the age range, we have plotted Figure 2 for males and Figure 3 for females. (a) the plot represents the RMSE for ages between 0 and 60; meanwhile, (b) the plot shows above 60 (the age range more interesting for insurance companies).

Next, if we show the models by age, there are not many differences between mortality models in females. Meanwhile, in male populations, we can see that the

additive model works better for all ages. Although, it should be remarked that differences are higher in old ages, 60 to 90. This fact is interesting because this range of ages has a strong interest for the insurance companies.

Table 1. RMSE obtained by LOOCV in each studied region for all ages.

Measure	RMSE			
Gender	Male		Female	
Model/Region	Additive	Multiplicative	Additive	Multiplicative
España	0.0054	**0.0052**	**0.0050**	0.0052
Andalucía	**0.0066**	0.0071	**0.0063**	0.0087
Aragón	**0.0084**	0.0084	**0.0067**	0.0070
Asturias	**0.0080**	0.0090	0.0059	**0.0057**
Baleares	0.0084	**0.0080**	**0.0056**	0.0064
Canarias	**0.0179**	0.0193	**0.0109**	0.0149
Cantabria	0.0089	**0.0086**	**0.0072**	0.0073
CLM	**0.0092**	0.0105	**0.0075**	0.0075
CYL	**0.0070**	0.0084	**0.0062**	0.0083
Cataluña	0.0062	**0.0060**	**0.0058**	0.0059
C. Valenciana	**0.0066**	0.0070	**0.0053**	0.0075
Extremadura	0.0095	**0.0091**	**0.0072**	0.0072
Galicia	**0.0077**	0.0080	0.0045	**0.0038**
Madrid	0.0094	**0.0081**	0.0071	**0.0069**
Murcia	**0.0085**	0.0086	**0.0078**	0.0089
Navarra	**0.0095**	0.0104	**0.0076**	0.0096
País Vasco	0.0070	**0.0067**	**0.0061**	0.0066
La Rioja	**0.0107**	0.0114	**0.0084**	0.0093
TOTAL	**0.0097**	0.0101	0.0489	**0.0488**

5. CONCLUSIONS

This paper has shown how a specific kind of resampling method, LOOCV, is a proper technique to decide which multi-population mortality model produces the best forecasting accuracy.

This paper compares two multi-population mortality models using LOOCV and shows which model produces the best forecasting outcomes. The outcomes have shown that there are few differences between the models. Although, in the males, the measure of accuracy is better for the additive mortality model, and the precision is also better for all ages, which is interesting for insurance companies.

Figure 2. RMSE for additive and multiplicative models along ages in all studied regions for males; (a) ages 0-60 and (b) ages 65-90.

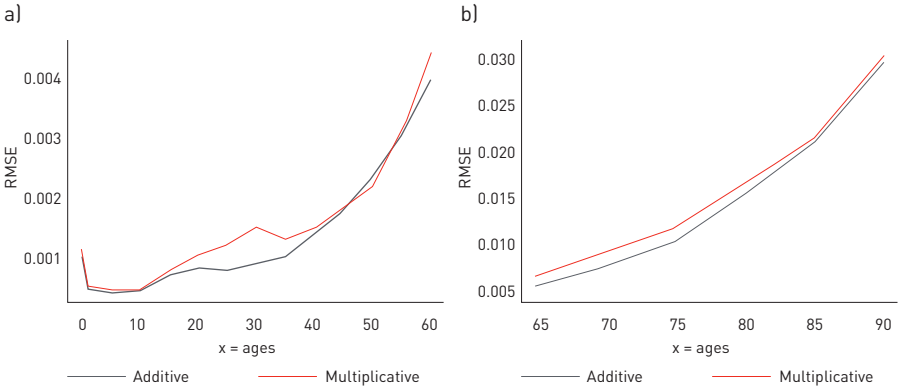
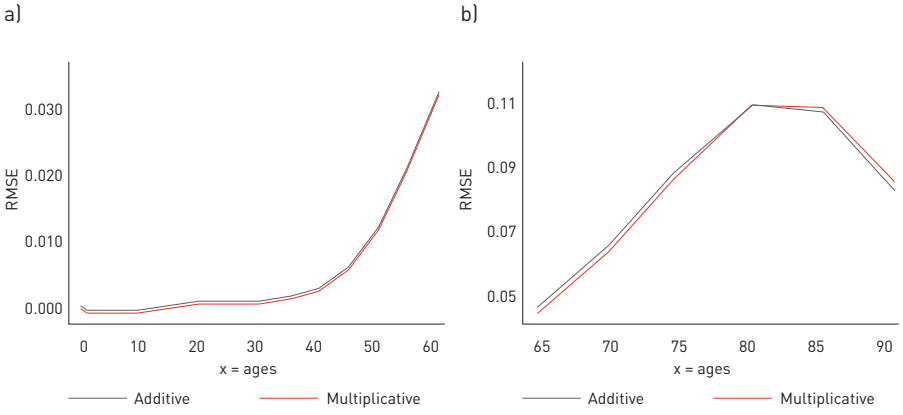


Figure 3. RMSE for additive and multiplicative models along ages in all studied regions for females; (a) ages 0-60 and (b) ages 65-90.



REFERENCES

Atance, D., Debón, A., and Navarro, E. (2020). "A comparison of forecasting mortality models using resampling methods". Mathematics, 8, 9, 1550.

Debón, A., Montes, F., and Martínez-Ruiz, F. (2011). "Statistical methods to compare mortality for a group with non-divergent populations: an application to Spanish regions". European Actuarial Journal, 1, 2, 291-308.

Dong, Y., Huang, F., Yu, H., and Haberman, S. (2020). "Multi-population mortality forecasting using tensor decomposition". Scandinavian Actuarial Journal, 2020, 8, 754-775.

Gareth, J., Daniela, W., Trevor, H., and Robert, T. (2013). "An introduction to statistical learning: with applications in R". Springer.

Lee, R.D., and Carter, L.R. (1992). "Modeling and forecasting US mortality". Journal of the American Statistical Association, 87, 419, 659-671.

Li, N., and Lee, R. (2005). "Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method". Demography, 42, 3, 575-594.

R Core Team (2022). "R: a language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria.

Russolillo, M., Giordano, G., and Haberman, S. (2011). "Extending the Lee-Carter model: a three-way decomposition". Scandinavian Actuarial Journal, 2011, 2, 96-117.

AUTOMOBILE PASSENGER INJURIES ACCORDING TO AGE AND CRASH LOCATION

Luis E. Céspedes
Zurich Insurance, Spain
lcespeco7@alumnes.ub.edu

Mercedes Ayuso
Universitat de Barcelona, Department of Econometrics-Riskcenter-IREA, Spain mayuso@ub.edu

ABSTRACT

Introduction: We analyze some potential determinants of car passenger injuries, with a focus on occupants aged 65 and more years. We pay special attention to the geographical location of the crash, with an interest on the comparison between rural and urban areas. Method: Drawing from data from the Spanish Traffic Authority for crashes in 2016, we use a multinomial regression logistic model to investigate the effects of a set potential risk factors on different severity bodily injury damages for passengers in a vehicle. Results: Passengers over 75 years old are significantly more likely to die than passengers below 65 years old when involved in a crash. Besides crashes in towns and rural areas are less likely to produce non-serious injuries in passengers compared to those that took place in cities; the reverse relationship holds for serious injuries. By gender, women passengers are more likely to suffer non-serious injuries.

1. INTRODUCTION

Spain is the 2nd largest country of the European Union by area, with a total of 505.983 km2 (INE, 2022), and the 17th when it comes to inhabitants by km2, below

all of its European neighbors (France 14th, Italy 6th and Portugal 10th). In the European context, Spain combines the lowest density of settlements (almost 90% of its territory is uninhabited) and areas with the highest population concentration, a highly anomalous pattern that emerged from its history and that cannot be accounted by climatic conditions, as it is the case for Scandinavian countries (Gutiérrez et al., 2021).

The population distribution has an impact on economic activity as, with industrialization and structural change, when people and firms locate close to each other in cities and industrial clusters a benefit arise, named agglomeration economies (Glaeser, 2010). These benefits are conditioned by returns to scale and transportation costs, which affect the relationship between economics agents (suppliers and firms, employers and workers, and producers and consumers). The spatial concentration of economic activity increases market access, resulting in cheaper and more varied inputs, as well as allows the sharing of risk and indivisible facilities, such as hospitals or universities (Beltrán et al., 2018).

The characteristics of rural areas, dispersed and low density, create difficulties to provide equal access to the same public services and resources offered to city-dwellers, and condition the well-being of these communities (Camarero and Oliva, 2019). As note by Camarero and Oliva, the 22% of households have some or great difficulties to access primary healthcare services, while in cities only a 7,4% do so, and regarding public transportation, the 21,7% of rural households do present difficulties to access it while in cities only a 4,3% do. Therefore, the lack of some public services and an insufficient public transport network, make rural-dwellers more reliant on their private automobiles to satiate their big mobility needs.

Old people use more frequently healthcare services and are key to their well-being, but those in rural areas face inequality in the access to healthcare facilities, mainly located in medium and high-density population areas, in comparison to elder city-dwellers. Therefore, those in rural areas have to drive to far away municipalities by car to access these services, with the help of someone else or on their own. Therefore, both drivers and passengers from thinly populated areas might be older than the general average. The effects of longevity are present in

their characterization, as joint life expectancy of couples, i.e. the number of years a couple lives until one of them dies, is ever-increasing (Alaminos and Ayuso, 2016).

There is an institutional demand to identify the effects of these new mobility patterns emerging from the generational shift to older societies (INE, 2022), especially on traffic crashes with vehicle occupants of advanced ages (Alemany et al., 2013). One of them is the United Nations (UN General Assembly, 2015) with their "2030 Agenda", which has put an emphasis on promoting actions to achieve Sustainable Development Goals (SDG) on key areas. Some relevant SDGs related to the mobility issues are the need to engage in actions to assure good health and well-being (SGD 3), to invest in infrastructure (SDG 9), to reduce inequalities (SDG 10) and to create sustainable cities and communities (SDG 11). The World Health Organization has stressed the need to adapt the implementation of actions in the SDG framework to empower older people, to help them maintain their functional abilities (WHO, 2020), and has put great emphasis on their need to be able to meet their own basic needs and to be mobile.

Researchers have started to focus on these issues, as illustrated by the creation of a composite indicator centered around old people and SDG completion (Alemany et al., 2021). These research lines are useful to center the public debate around topics that lack visibility but whose outcomes have a relevant impact on society. Mobility between areas with different population densities is one of them. It is an important topic for public and private stakeholders, as profiling their vehicle occupants helps to understand better their motives to drive and to shape urban planning, to contribute to understand crashes which are, then, accounted and forecasted by public and private health care systems, insurance companies, governmental traffic agencies and even vehicle manufacturers, with aims ranging from implementing policies to reduce accidents or accident severity, the calculation of more adjusted provisions (Ayuso et al., 2020), reevaluation of compensations in case of accidents or even improving car designs to make them safer (Ayuso et al., 2019; Guillen et al., 2019).

Elder drivers are aware of their limitations and some self-regulate the number of kilometers they drive, for instance many may avoid driving at night in the rain. However, their capacity to self-regulate is limited by their desire to keep their

lifestyle, the unavailability of family and friends to provide transport when needed or unwillingness to ask them for help with transportation, and the unavailability of public transportation (Baldock et al., 2006). So, it is of concern that their actual fatality rates are higher than observed, as there are few elder drivers and some self-regulate their exposure to risk of crash (Rolison et al., 2012).

In this work we seek to understand the relationship between driver ages and the location of the accident on the severity of passenger injuries (non-serious, serious or fatal). To do so, we examine automobile crashes that took place in Spain in 2016, segmenting the analysis according to the location of the accident (densely populated area, intermediate populated areas or rural areas) and to the age of the driver (under 65 years, between 65 and 75 years, and over 75 years old). In this paper we analyze the data for vehicle occupants, emphasizing different perspectives and segmentations for the ages of drivers and passengers, with plots and a correlation analysis. After that, we describe the data used to model passenger injuries. Finally, we present the multinomial logistic regression results and remark some important outcomes to conclude.

2. DRIVER CENSUS

In January of 2016, there were a total of 26,5 million registered drivers in the Spanish driver census. Over an 85,1% of them were between the ages of 15 and 64, the 9,8% were between 65 and 74 years, while the remaining 5,2% were 75 or more years old. As for the urbanization of their municipality of residence, the 52,6% lived in cities, the 33,4% in towns and suburbs and a 14% dwelled in rural areas.[1]

Drivers of the age groups 15-64 and 65-74 displayed similar patterns. Their distribution across the different geographical areas was similar, as was their share of total drivers in each of these areas. Only drivers older than 74 years showed significant differences. It was the group of individuals with the highest concentration

[1] This geographical segmentation is based on the Eurostat classification methodology, according to population size and density, from the most populated to the least (cities, towns and suburbs, and rural areas).

in rural areas, over a 21,9% of them dwelled there in comparison to the 13,4% and 15,1% of drivers with ages 15-64 and 65-74.

3. ESTIMATION OF A LOGISTIC REGRESSION MODEL

We wanted to evaluate the significance of potential risk factors on each level of passenger injuries severity. Each injury variable was binary, the passenger could either be injured or not, and only one injury kind could be experience. With all things considered, the logistic regression positioned itself as a good modelling choice, as it allowed to estimate probabilities of injuries with a dependent binary variable and it offered the direction of association.

$$P(Y_i = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)} \; where\ i = injury\ type, 1\ is\ Yes$$

$$P(Y_i = 0) = \frac{1}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)} \; where\ i = injury\ type, 1\ is\ Yes$$

4. MODELLED VARIABLES

We aimed at modelling the probability of a passenger from a passenger car with only 2 occupants (the driver and the passenger) of suffering an injury (non-serious, serious or fatal) in a crash with another passenger car. To do so, we included thirteen variables in the logit regressions, all of them presented in Table 1. The number of observations we used was enough to be representative at a 99% confidence interval and a 2% error, a total of 4.868 observations with complete information.

The crashes took place mostly in densely (54,8%) and intermediate population density areas (32,5%), evenly split between taking place in and outside intersections, and mostly in streets (50,8%), secondary roads (29,4%) and highways (15,4%).

The drivers of this data set were mostly male (68,6%), they were overwhelmingly below 65 years old (89,2%) and were considered as responsible for the crash by the police in a 58,8% of the cases.

The passengers were mostly female (66,2%), younger than 65 years old (88,2%), they were seated in the front row (87,2%) and wore the seat belt (72,1%). More than a quarter of passengers suffered some kind of injury: non-serious injuries were experienced by a 25,2% of passengers, serious injuries by a 1,6% and, finally, fatal injuries by a 0,5%.

Table 1. Frequencies of modelled variables

Variable	Level	Abs. Freq.	Rel. Freq.
Driver Gender	Male	3.341	68,6%
	Female	1.527	31,4%
Driver Crash responsability	Yes	2.862	58,8%
	No	2.006	41,2%
Passenger gender	Male	1.647	33,8%
	Female	3.221	66,2%
Passenger used seat belt	Yes	3.512	72,1%
	No	109	2,2%
	Unknown	1.247	25,6%
Passenger vehicle location	Front seats	4.247	87,2%
	Back seats	533	10,9%
	Other seats	88	1,8%
Road type	Highways	749	15,4%
	Secundary roads	1.542	31,7%
	Street	2.471	50,8%
	Other	196	4,0%
Intersection	Inside one	2.461	50,6%
	Ountside one	2.407	49,4%

Variable	Level	Abs. Freq.	Rel. Freq.
Crash location urbanization	Cities	2.667	54,8
	Towns and suburbs	1.583	32,5
	Rural area	618	12,7
Driver Age Bracket	< 65	4.341	89,2
	65-75	368	7,6
	> 75	159	3,3
Passeger Age Bracket	< 65	4.295	88,2
	65-75	400	8,2
	> 75	173	3,6
Non-serious injuries	Yes	1.229	25,2
	No	3.639	74,8
Serious	Yes	78	1,6
injuries	No	4.790	98,4
Fatal injuries	Yes	24	0,5
	No	4.844	99,5

6. RESULTS

Table 2 contains the estimated coefficients for the logit regressions for the three injury levels. The logit regression evaluates whether the presence of a level different to the base one has an impact on the probability of occurrence of the dependent variable. As such, when a coefficient for a variable level is positive and significant, it means that the presence of that level increases the odds of suffering the studied injury in comparison to the base level. On the contrary, a negative significant value implies that the odds of occurrence of the injury type declines.

No variable is significant in all the three models, but the driver's gender and the variable that indicates whether the crash occurred in an intersection. In a crash, when the driver is a female, the passenger is less likely to suffer any kind of injury. Conversely, if the accident takes place outside an intersection, the passenger is more likely to suffer any injury in comparison to those that take place in intersections.

The age of the driver was not a key variable, as it only reduced the odds of non-serious injuries when drivers were aged 65-75. As for the responsibility of the crash, when the driver was responsible, its passenger was less likely to suffer non-serious injuries and more likely to suffer serious ones.

Table 2. Logit regressions result by passenger injury severity level.

Variables		Logit models					
		Non-serious inuries		Serious injuries		Fatalities	
		Coef.	P-Value	Coef.	P-Value	Coef.	P-Value
Driver's gender							
	Intercept	0,8883	0,00143 ***	-5,4746	0,0000 ***	-4,1489	0,00073 ***
	Man	–	–	–	–	–	–
	Woman	-0,2453	0,0016 ***	-0,5030	0,0986 *	-1,5538	0,0306 **
Driver crash responsability							
	No	–	–	–	–	–	–
	Yes	-1,1454	0,0000 ***	0,5795	0,0142 **	0,7202	0,1033
Passenger gender							
	Man	–	–	–	–	–	–
	Woman	0,5548	0,0000 ***	0,3116	0,2452	0,0355	0,9423
Passenger used seat belt							
	No	–	–	–	–	–	–
	Yes	-0,0523	0,8341	-1,5919	0,0024 ***	-2,2065	0,0101 **
	Unknown	-0,0633	0,8049	-2,5733	0,0001 ***	-1,7846	0,0810 *
Road type							
	Highway	–	–	–	–	–	–
	Secondary roads	0,3016	0,0042 ***	1,9386	0,0003 ***	0,6741	0,2157
	Street	1,0075	0,0000 ***	0,9027	0,1255	-24411	0,0354 **
	Other	0,5325	0,0048 ***	0,6482	0,4622	-15,4795	0,9890
Intersection							
	In an intersection	–	–	–	–	–	–
	Outside an intersection	0,2297	0,0015 ***	0,6920	0,0046 ***	1,4169	0,0072 ***

Variables		Logit models					
		Non-serious inuries		Serious injuries		Fatalities	
		Coef.	P-Value	Coef.	P-Value	Coef.	P-Value
Crash geographical location degreen of urbanization							
	Densely populated area	–	–	–	–	–	–
	Intermediate density area	-0,2687	0,0014 ***	0,5865	0,0792 *	0,9176	0,1469
	Thinly populated area	-0,2328	0,0470 **	1,0114	0,0056 ***	1,0111	0,1288
Driver age bracket							
	< 65	–	–	–	–	–	–
	65-75	-0,3655	0,0125 **	0,2225	0,5997	0,1207	0,8828
	> 75	0,1077	0,6227	-0,3493	0,5558	0,8308	0,2722
Passenger age bracket							
	< 65	–	–	–	–	–	–
	65-75	-0,0628	0,6586	0,2349	0,5836	0,3588	0,6431
	> 75	-0,0439	0,7950	0,6592	0,1834	1,6194	0,0173 **

* P-Value < 0,1; ** P-Value < 0,05; *** P-Value < 0,01

The gender of the passenger was only significant for non-serious injuries, where being a woman implied higher odds of experiencing them. We were not surprised to find that passengers that wore a seat belt were less likely to suffer serious or fatal injuries in comparison to those who did not put on it. The location of the passenger inside the car mattered for the odds of fatal injuries, as those in front row seats were less likely to die than those in back row seats. As for their age, passengers older than 75 years old had a higher probability of dying when involved in a crash.

Finally, focusing on the geographical location, when the crash takes place in thinly or intermediate population density areas, the passengers are less likely to suffer non-serious injuries and more likely to experience serious ones. As for the road type, the odds of non-serious injuries increase when the accident occurs in secondary roads, streets or other kind of roads rather than in highways. Crashes in secondary roads are also more likely to cause serious injuries in the passenger. As for streets, in comparison to highways, they reduce the probability of fatal injuries.

7. CONCLUSIONS

Our results contribute to certify the inequality of outcomes for crashes depending on the geographical location of the accident. Traffic crashes in towns and rural areas are more prone to have worse outcomes for the passengers involved. Under the umbrella of SGD 9 to 11, policy makers need to adapt infrastructure when needed and propose creative solutions to make them safer, to reduce these inequalities and make these communities more sustainable.

As for SGD 3, good health and well-being, the analysis of passenger injuries remarks that age does matter, as when they are over 75 years of age, they are more likely to experience fatal injuries, possibly due to the fragility of their bodies. We also demonstrate that old drivers tend to drive with people their age, and most passengers tend to be female, so we could expect that if there are more older driver crashes, there is likely to be an increase of deaths of old women.

Further research is needed to understand better the characteristics of crashes according to the geographical location and the typology of accidents where elder drivers are involved, as to isolate the factors that cause such increases in the severity of injuries.

Finally, we have mentioned several areas where there are inequalities and people experience worse outcomes depending on their age, on the place the crash takes place, on the sex of the driver, etc. Reducing these inequalities requires the collaboration of society as a whole, as SDG 17 remarks, to improve the quality of life of elder people and allow them to fulfill their needs. If not, society will not be ready to face the challenges of an ageing society, with consequences that will start to be seen no further than the next decade.

ACKNOWLEDGMENTS

We want to acknowledge the support received by the Spanish Traffic Authority (DGT) by providing us with the data sets to perform this analysis. Mercedes Ayuso is grateful to the Spanish Ministry of Science and Innovation for the grant

PID2019-105986GB-C21, and to the Generalitat de Catalunya for the grant 2020-PANDE-00074.

REFERENCES

Alaminos, E., and Ayuso, M. (2016). "Modelo actuarial multiestado para el cálculo de probabilidades de supervivencia y fallecimiento según estado civil: Una aplicación al pago de pensiones concurrentes". Anales del Instituto de Actuarios españoles, 22, 41-71.

Alemany, R., Ayuso, M., and Céspedes, L. (2021). "Indicador de Calidad de Vida de los adultos mayores en España y en Europa". Indicador ODS-IVDS65+. RISKCenter, Cátedra UB-Escuela de Pensamiento Fundación Mutualidad de la Abogacía sobre Economía del Envejecimiento. https://www.escueladepensamiento.org/wp-content/uploads/2021/11/Publicacion_fin

Alemany, R., Ayuso, M., and Guillen, M. (2013). "Impact of road traffic injuries on disability rates and long-term care costs in Spain". Accident Analysis and Prevention, 60, 95-102.

Ayuso, M., Guillen, M., and Nielsen, J.P. (2019). "Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data". Transportation, 46, 3, 735-752.

Ayuso, M., Sánchez, R., and Santolino, M. (2020). "Does longevity impact the severity of traffic crashes? A comparative study of young-older and old-older drivers". Journal of Safety Research, 73, 37-46.

Baldock, M.R.J., Mathias, J.L., McLean, A.J., and Berndt, A. (2006). "Self-regulation of driving and its relationship to driving ability among older adults". Accident Analysis & Prevention, 38, 5, 1038-1045.

Beltrán Tapia, F., Díez-Minguela, A., and Martinez-Galarraga, J. (2018). "Tracing the Evolution of Agglomeration Economies: Spain, 1860–1991". The Journal of Economic History, 78, 1, 81-117.

Camarero, L., and Oliva, J. (2019). "Thinking in rural gap: mobility and social inequalities". Palgrave Communications, 5, 1. https://doi.org/10.1057/s41599-019-0306-x

Glaeser, E. L. (2010). "Introduction" In Agglomeration Economics, edited by Edward L. Glaeser, 1-14 Chicago: University of Chicago.

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

Gutiérrez, E., Moral-Benito, E., Oto-Peralías, D., and Ramos, R. (2020). "The spatial distribution of population in Spain: an anomaly in European perspective". Documentos de Trabajo, n.º 2028, Banco de España.

INE (2022). "Territorio y medio ambiente". España en cifras 2022. ISSN: 1136-1611.

INE (2022). "Población. España en cifras 2022". ISSN: 1136-1611.

Rolison, J.J., Hewson, P.J., Hellier, E., and Husband, P. (2012). "Risk of fatal injury in older adult drivers, passengers, and pedestrians". Journal of the American Geriatrics Society, 60, 8, 1504-1508.

UN General Assembly (2015). "Resolution Adopted by the General Assembly on 25 September 2015. Transforming Our World: The 2030 Agenda for Sustainable Development".

World Health Organization (2020). "Decade of healthy ageing: baseline report". ISBN 978-92-4-001791-7.

HYPOTHESIS TESTS AND BAYESIAN METHODS TO DEAL WITH OFF-DIAGONAL ELEMENTS IN CONFUSION MATRICES

Inmaculada Barranco-Chamorro
Universidad de Sevilla, Facultad de Matemáticas, Departamento de Estadística e I.O., España
chamorro@us.es

Rosa M. Carrillo-García
Universidad de Sevilla, Facultad de Matemáticas, Departamento de Estadística e I.O., España

ABSTRACT

Tests to deal with the off diagonal elements in confusion matrices are proposed. They are tailored to detect problems of bias of classification among classes. A Bayesian approach is developed aiming to estimate overprediction and underprediction probabilities among classes.

1. INTRODUCCIÓN

Confusion matrices are the standard way of summarizing the performance of a classifier. It is assumed that the qualitative response to be predicted has r >2 categories, the confusion matrix will be a r x r matrix, where the rows represent the actual or reference classes and the columns the predicted classes (or vice versa). So the diagonal elements correspond to the items properly classified, and the off-diagonal to the wrong ones. Most papers dealing with confusion matrices focus on the assessment of the overall accuracy of the classification process, such as kappa coefficient, and methods to improve these measurements, see for instance Grandini et al. (2020) and references therein. However, a scarce number of papers consider the study of the off-diagonal cells in a confusion matrix. In this

paper a method is proposed that can be useful for a better definition of classes and to improve the global process of classification.

Based on the results given in Barranco-Chamorro and Carrillo-García (2021), the problem of classification bias is introduced. This is a kind of systematic error, which may happen between categories in a specific direction. If a classifier is fair or unbiased, then the errors of classification between two given categories A and B must happen randomly, that is, it is expected that they occur approximately with the same relative frequency in every direction. Quite often, this is not the case, and a kind of systematic error or bias occurs in a given direction.

The classification bias can be due to deficiencies in the method of classification. For instance, it is well known, Goin (1984), that an inappropriate choice of k in the k-nearest neighbor (k-nn) classifier may produce this effect. In case of being detected, the method of selection of k must be revised. On the other hand, the classification bias may be caused by the existence of a unidirectional confusion between two or more categories, that is, the classes under consideration are not well separated. Anyway, in case of being detected this problem, the process of classification should be improved. To identify this problem in a global way, marginal homogeneity tests are proposed. The tests are based on Stuart--Maxwell test, (Black and Gonen, 1997), and Bhapkar test, (Sun and Yang, 2009). If the null hypothesis of marginal homogeneity is rejected, a One versus All methodology is proposed, in which Mc-Nemar type tests, (McNemar, 1947), are applied to every pair of classes. Second a Bayesian method based on the Dirichlet-Multinomial distribution is developed to estimate the probabilities of confusion between the classes previously detected. So it can be assessed in a formal way, if certain classes suffer from a problem of overprediction or underprediction. To illustrate the use of our proposal, real applications are considered. They are taken from the fields of Geostatistics and Bioinformatics. As computational tools, we highlight that the R Software and R packages are used.

2. CONCLUSIONS

In this paper, methods to detect the bias of classification, as well as overprediction and underprediction problems associated to categories in a confusion matrix are proposed.

They may be applied to results of applying supervised learning algorithms, such as logistic regression, linear and quadratic discriminant analysis, naive Bayes, k-nearest neighbors, classification trees, random forests, boosting or support vector machines, among others.

Marginal homogeneity tests for matched pairs of observations are proposed.

Also the Multinomial-Dirichlet distribution is applied to asses the probabilities of over- and under-prediction in a misclassification problem. Two applications taken from peer-reviewed and different scientific areas have been carried out. The results are satisfactory.

We highlight that our proposal is of interest for a better definition of classes, and to improve the performance of classification methods.

ACKNOWLEDGMENTS

The research of Rosa M. Carrillo-García has been funded by Grant PI3 "Programa IMUS de Iniciación a la Investigación", IMUS, Seville, 2021.

REFERENCES

Barranco-Chamorro I., and Carrillo-García R.M. (2021). "Techniques to deal with off-diagonal elements in confusion matrices". Mathematics, 9, 244, 3233.

Barranco-Chamorro, I., Luque-Calvo, P., Jiménez-Gamero, M., and Alba-Fernández, M. (2017). "A study of risks of Bayes estimators in the generalized

half-logistic distribution for progressively type-II censored samples". Mathematics and Computers in Simulation., 137, 130-147.

Black, S., and Gonen, M. (1997). "A generalization of the Stuart-Maxwell test". In SAS Conference Proceedings: South-Central SAS Users Group 1997; Applied Logic Associates, Inc.: Houston, TX, USA.

Franco M., and Vivo J.M. (2021). "Evaluating the performances of biomarkers over a restricted domain of high sensitivity". Mathematics, 9, 21, 2826.

Goin, J.E. (1984). "Classification bias of the k-nearest neighbor algorithm". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-6, 379-381.

Grandini, M., Bagli, E., and Visani, G. (2020) "Metrics for multi-class classification: An overview". arXiv 2020, {arXiv:2008.05756}.

McNemar, Q. (1947). "Note on the sampling error of the difference between correlated proportions or percentages". Psychometrika, 12, 153-157.

Pérez, C.J., Girón, F.J., Martín, J., Ruiz, M., and Rojano, C. (2007). "Misclassified multinomial data: A Bayesian approach". Revista de la Real Academia de Ciencias. Exactas, Físicas y Naturales. Serie A Matemáticas. (RACSAM), 101, 71-80.

Sun, X., and Yang, Z. (2008). "Generalized McNemar's test for homogeneity of the marginal distributions". In Proceedings of the SAS Global Forum Proceedings. Statistics and Data Analysis, San Antonio, TX, USA, 16-19 March. 382, 1-10.

QUANTILE CAPITAL ALLOCATION AND DEPENDENCE STRUCTURE OF RISKS

Jaume Belles-Sampera
Universitat de Barcelona, Catalana-Occidente, RISKcenter, España,
belles.sampera@ub.edu,

Miguel Santolino
Universitat de Barcelona, Dept.Econometria, Estadística y E.A., RISKcenter, España
msantolino@ub.edu

ABSTRACT

Capital allocation problems arise when a total risk amount associated with the aggregate losses has to be distributed across the multiple units of risk that make up these total risk losses. The risk capital allocation problem can be interpreted as on optimization problem in which the weighted sum of measures for the deviations of the risk unit's losses from their respective allocated capitals is minimized (Dhaene et al., 2012). Given that the total amount to allocate is computed as the α-quantile of the sum of risks, the quantile allocation principle allocates to each risk the same probability level p-quantile amount such that the full allocation requirement is satisfied. This means that a constant proportional reduction (or increase) is applied on the probability level of the individual risk quantiles, i.e. $p=f\cdot\alpha$ being $f>0$ the multiplier to reduce (or increase) α. Dhaene et al. (2003), and later generalized by Cai and Wang (2021), showed that the p-quantile allocation rule derived from the sum of comonotic risks is the solution of the optimization problem based on the absolute deviation criterion. In this article we discuss the key role that the dependence structure plays on the distance between α and p probability levels, given the marginal distributions of the loss variables. The scale of the loss variables has also impact on the distance between the probability levels, particularly when one random loss is much larger than the others. A set of

examples is provided to illustrate how the distance between α and p probability levels is affected by the dependence structure and scale of random variables. To summarize, the functional form of *f* in the α-range describes the multivariate distribution.

1. INTRODUCCIÓN

Risk in financial and actuarial applications is often defined as a random variable associated with costs or losses. Capital allocation problems arise when a total amount associated with the aggregate risk has to be distributed across the multiple units of risk that make up this total risk. Examples of capital allocation problems can be found, for instance, in asset allocation strategies for portfolio selection, the allocation of the total solvency capital requirement across business lines or when distributing total claims administration costs among policies in the portfolio, among others.

The capital allocation principle is the set of guidelines that indicates how the total amount has to be allocated. There is an extensive amount of capital allocation principles proposed in the literature. Some capital allocation principles have been motivated based on game theory in which capital allocation problems are interpreted as coalition games (Denault, 2001; Tsanakas, 2009). An alternative approach to derive capital allocation principles is from the economy theory, as optimization problems in which a loss function of particular interest for risk managers is minimized (Furman and Zitikis, 2008; Dhaene et al., 2003; Zaks et al., 2006; van Gulick et al., 2012).

Dhaene et al. (2012) proposed a general theoretical framework in which a capital allocation principle is the outcome of a particular optimization problem. Many of the existing capital allocation principles used in practice can be accommodated in the framework, including the haircut allocation principle (Belles-Sampera et al., 2022). The optimization criterion involving the use of a quadratic function to compute deviations of the outcomes of the losses from their allocated capital has received most attention in literature (Belles-Sampera et~al., 2022; Zaks and Tsanakas, 2014). The absolute deviation function has received less attention. [Dhaene et al., 2003), and later generalized by Cai and Wang (2021), showed that

the quantile allocation criterion is the solution of the optimization problem based on the absolute deviation function.

We here focus on the quantile allocation principle. Dhaene et al. (2012) argued that, whether the total amount to share is considered exogenous, the allocated capitals to each risk are not influenced by the dependence structure between the different losses when the quantile capital allocation principle is applied. In this article we analyse the case in which total amount to share is not exogenous. When the total amount to share is computed as the α-quantile of the aggregate losses, we show that the dependence between individual risks plays a key role in the p-level quantile allocation solution. We argue that the distance between α and p probability levels depends on the dependence structure of the loss variables and also in the shape and scale of the marginal distributions. It is shown how the quantile allocation principle can be used to describe the aggregating behaviour of the multivariate distribution involved. It can be used, for instance, to unveil α values for which the Value-at-Risk of the sum of the marginals is super or subadditive.

The article is structured as follows. The general optimal capital allocation framework is defined in the next section. Section (3) analyses the quantile allocation principle as the solution to absolute deviation optimization problem. Examples of the impact of the dependence structure and the shape and scale of variables are provided in Section (4). Section (5) concludes.

2. OPTIMAL CAPITAL ALLOCATION AS THE SOLUTION OF A MINIMIZATION PROBLEM

Assume that a capital $K > 0$ has to be allocated across J business units denoted by $j = 1,..., J$. According to Dhaene et al. (2012), most capital allocation problems can be described as the optimization problem given by[2]

[2] The original allocation problem proposed by [Dhaene et~al., 2012] was with the characteristics (b) and (c) in (1) as follows,

- a set of non-negative weights v_j, $j = 1,..., J$, such that $\sum_{j=1}^{J} \boldsymbol{v_i} = \mathbf{1}$; and
- a set of non-negative random variables ζ_j, $j = 1,..., J$, with $\mathbb{E}[\zeta_j] = 1$. We here follow the generalization of Cai and Wang (2021)

$$\min_{K_1,K_2,\ldots,K_J} \sum_{j=1}^{J} v_j \mathbb{E}\left[\zeta_j D\left(\frac{X_j - K_j}{v_j}\right)\right] \quad s.t. \quad \sum_{j=1}^{J} K_j = K, \qquad (1)$$

with the following characterizing elements:

- a function D: $\mathbb{R} \to \mathbb{R}^+$;
- a set of positive real numbers υ_j, $J = 1,\ldots, J$; and
- a set of positive random variables ζ_j such that $\mathbb{E}[\zeta_j] > 0$, $j = 1,\ldots, J$.

If $D(x) = x^2$ is selected then the optimization criterion linked to (1) is called a *quadratic optimization criterion*. The optimal capital allocation under this setting is provided in Dhaene et al. (2012). An alternative proof of the solution with quadratic optimization criterion can be consulted in Belles-Sampera et al. (2022). In this article we deal with the absolute deviation function, $D(x) = |x|$, where it is differentiated between surplus and shortfall risks $|x| = x_+ + x_-$ being $x_+ = \mathbf{max}\{x, 0\}$ and $\mathbf{max}\{-x, 0\}$.

3. OPTIMAL QUANTILE CAPITAL ALLOCATION

Let us define the following optimization problem in the framework (1) based on the absolute deviation function and with $\upsilon_j = 1$, as follows

$$\begin{cases} \min \sum_{j=1}^{J} \left(\mathbb{E}\left[\zeta_j (X_j - K_j)_+\right] + \mathbb{E}\left[\psi_j (X_j - K_j)_-\right]\right) \\ s.t. \quad \sum_{j=1}^{J} K_j = K \end{cases} \qquad (2)$$

Cai and Wang (2021,Theorem 4.2) showed that (2) has the following solution

$$K_j^* = F_j^{-1(p_j)}\left(\frac{\mathbb{E}[\zeta_j] - \lambda}{\mathbb{E}[\zeta_j + \psi_j]}\right), \quad j = 1, \ldots, J \qquad (3)$$

with $p_j \in [0,1]$, and λ, p_j are such that

$$\begin{cases} \sum_{j=1}^{J} F_j^{-1(p_j)}\left(\frac{\mathbb{E}[\zeta_j - \lambda]}{\mathbb{E}[\zeta_j + \psi_j]}\right) = K \\ \underline{M} < -\lambda < \overline{M} \end{cases}$$

where $F_j^{-1(p_j)}(\alpha) = p_j F_j^{-1}(\alpha) + (1-p_j)F_j^{-1+}(\alpha)$, with $F_j^{-1}(\alpha)$ and $F_j^{-1+}(\alpha)$ being the left-continuous and right continuous inverse functions of the distribution function of X_j, $F_j(x) = Pr\{X_j \le x\}$, $j = 1,\dots,J$, which are defined as

$$F_j^{-1}(\alpha) = \inf\{x \in \mathbb{R}, F_j(x) \ge \alpha\} = \sup\{x \in \mathbb{R}, F(x) < \alpha\}$$

$$F_j^{-1+}(\alpha) = \sup\{x \in \mathbb{R}, F_j(x) \le \alpha\} = \inf\{x \in \mathbb{R}, F(x) > \alpha\}$$

and $\underline{M}$ and $\overline{M}$ are defined as

$$\underline{M} = \max\{-\mathbb{E}[\zeta_j], j = 1,\dots, J\} < 0$$

$$\underline{\boldsymbol{M}} = \min\{-\mathbb{E}[\psi_j], j = 1,\dots, J\} > 0$$

Let assume now that F_j^{-1} , $j = 1,\dots, J$, are continuous and strictly increasing on $(0,1)$ and $\sum_{j=1}^{J} F_j^{-1}\left(\frac{\mathbb{E}[\zeta_j]-\underline{M}}{\mathbb{E}[\zeta_j+\psi_j]}\right) < K < \sum_{j=1}^{J} F_j^{-1}\left(\frac{\mathbb{E}[\zeta_j]-\overline{M}}{\mathbb{E}[\zeta_j+\psi_j]}\right)$, then the optimal is

$$K_j^* = F_j^{-1}\left(\frac{\mathbb{E}[\zeta_j]-\lambda}{\mathbb{E}[\zeta_j+\psi_j]}\right), \quad j = 1,\dots,J \tag{4}$$

with λ being the unique solution to $\sum_{j=1}^{J} F_j^{-1}\left(\frac{\mathbb{E}[\zeta_j]-\lambda}{\mathbb{E}[\zeta_j+\psi_j]}\right) = K$. .

Finally, let consider the case that $\mathbb{E}[\zeta_j]$, $j = \alpha$, $\mathbb{E}[\psi_j] = 1-\alpha$, $j = 1,\dots, J$, and $F_{S^c}^{-1+}(0) < K < F_{S^c}^{-1}(1)$, with is the comonotonic sum[3] of , then the solution of (3) is

$$K_j^* = F_j^{-1(p)}(F_{S^c}(K)), \quad j = 1,\dots,J \tag{5}$$

with $p \in [0,1]$ satisfying $F_{S^c}^{-1(p)}(F_{S^c}(K)) = K$. Note that (5) is the quantile capital allocation principle (Dhaene et al., 2012, Sec.2.3.2). In addition, if F_j^{-1}, $j = 1,\dots, J$, are continuous, then the optimal is

$$K_j^* = F_j^{-1}(F_{S^c}(K)), \quad j = 1,\dots,J. \tag{6}$$

[3] $S^c = \sum_{j=1}^{J} F_j^{-1}(U)$, where U is a uniform random variable

Remark: In the one hand, note that adapting expression (8) in [Dhaene et~al., 2012] to the notations used before, the quantile allocation principle can be expressed also as

$$\boldsymbol{K_j = F_j^{-1(p)}(f \cdot \alpha),\quad j = 1, \ldots, J,\quad s.t. \sum_{j=i}^{J} K_j = K.} \tag{7}$$

On the other hand, expression (3) when $\mathbb{E}[\zeta_j] = \alpha$, $\mathbb{E}[\psi_j] = 1 - \alpha$, $j = 1,\ldots, J$, and $\boldsymbol{F_{S^c}^{-1+}(0) < K < F_{S^c}^{-1}(1)}$ can be written as

$$\boldsymbol{K_j = F_j^{-1(p)}(\alpha - \lambda),\quad j = 1, \ldots, J,\quad s.t. \sum_{j=1}^{J} K_j = K.} \tag{8}$$

Therefore, the following relationship arises between f and λ:

$$\boldsymbol{f = 1 - \frac{\lambda}{\alpha}.} \tag{9}$$

4. APPLICATION

In order to investigate how the dependence structure and the shape and scale of random variables affect factors f (or λ) in quantile allocations, four different multivariate distributions have been selected and 99 of their respective *centile* allocations[4] have been calculated from 25000 simulations of each random vector. The four multivariate distributions are of dimension 3 and defined as follows:

[D1] The three marginals are normal random variables: V_1 with $\mu = 100000$ and $\sigma_1 = 1000$; $V_2 = -0.5 \cdot V_1 + N_a$, with N_a a normal rv with $\mu_a = 200000$ and $\sigma_a = 1000$; and $V_3 = V_1 + V_2 + N_b$, where N_b is another normal rv with $\mu_b = 100$ and $\sigma_b = 1000$. This is a multivariate distribution with a dependence structure not defined by an explicit copula but through relationships among its marginals. Note that $\mu_2 = 150000$, $\sigma_2 = 1000$, $\mu_3 = 250100$ and $\sigma_3 = 10000$.

[4] Quantile allocations for centiles α from $\alpha = 1/100$ to $\alpha = 99/100$.

[D2] The three marginals are normal random variables: V_1, V_2 and V_3 with μ_i and σ_i, i = 1, 2, 3 as for the first multivariate distribution. But the dependence structure among these marginals is set by a Clayton copula of dimension 3 with parameter θ = 5.25.

[D3] Three marginals equally distributed as Pareto α = 4, β = 1. The dependence structure is defined through a Gaussian copula with the following linear correlations among marginals: p_{12} = 0.7, p_{13} = –0.7 and p_{23} = –0.9.

[D4] This is a multivariate normal distribution: the marginals as in D2 and with linear correlations as in D3.

Figures 1 to 4 show the λ's and the f's obtained for each distribution. Horizontal lines for λ = 1 and f = 1 are plotted as a reference, because these values represents those quantiles for which the α-quantile of the sum of the marginals equals de sum of the α-quantiles of the marginals (strict additivity for VaR_a(S)). As it is illustrated in Table 1, if $\lambda < 0$ (or $f > 1$) then the sum of the marginals is super-additive and if $\lambda < 0$ (or $f > 1$) then the sum is sub-additive.

Distributions D1, D2 and D4 have marginals with the same shape and scale. As it is shown in Figures 1,2 and 4, the values of λ's and f's are really different, something that validates the intuition that the dependence structure plays and important role to determine those values. It is remarkable that the three distributions have sets of α confidence levels for which the VaR of the sum of the marginals cannot be sub-additive. In the case of D1 and D4, approximately, for $\alpha < 0.5$ super-additivity is observed while for $\alpha = >0.5$ sub-additivity is satisfied. This second statement is aligned, in the case of D4, with the result stated in McNeil et al. (2005, Theorem 6.8) about the sub-additivity of VaR for elliptical risk factors. For D2, the dependence structure seems that generate less volatility in the values of λ's for the centile allocations than the volatility observed for D1 and D4, although the set of confidence levels for which the sub-additivity is not satisfied is wider ($\alpha \in$ (0,0.8) approximately).

Figure 1. Values of λ's and f's for the alpha-quantile allocations of D1.

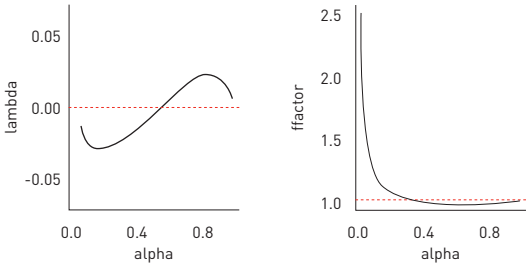


Figure 2. Values of λ's and f's for the alpha-quantile allocations of D2.

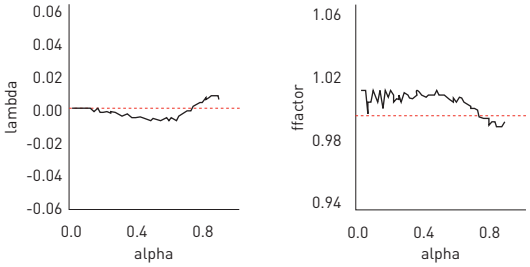


Figure 3. Values of λ's and f's for the alpha-quantile allocations of D3.

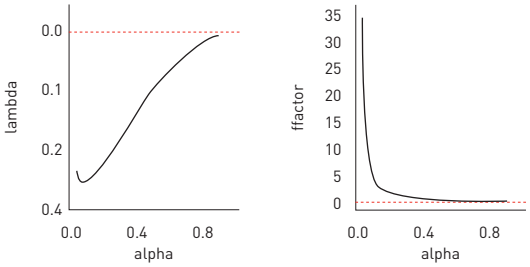
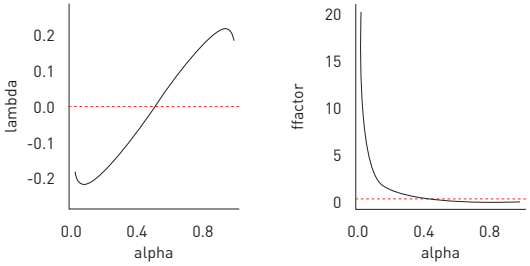


Figure 4. Values of λ's and f's for the alpha-quantile allocations of D4.



Distributions D3 and D4 share the same copula as dependence structure, but Figures 3 and 4 are completely different. Therefore, it seems that the shape and scale of the marginals is also playing a significant role in the values for quantile allocations, as it was expected. In fact, the values of the Pareto marginals in D3 are probably cancelling any potential effect of the Gaussian copula in the results. It is noticeable that we are facing and example for which sub-additivity can never be satisfied.

Table 1. Super- or sub-additivity of VaR for distributions D1 and D3. Bold-faced values indicate super-additivity.

Distribution	α	$VaR_\alpha(S)$	$VaR_\alpha(V_1)$	$VaR_\alpha(V_2)$	$VaR_\alpha(V_3)$	$VaR_\alpha(S) - \sum_{i=1}^{3} \boldsymbol{VaR}_\alpha(\boldsymbol{V_i})$
D1	0.4	497509.7	99737.51	149733.4	247532.3	**506.5714**
D1	0.8	508706.2	100841.1	150952.6	258567.6	-1655.191
D1	0.95	517097	101666.4	151836	266809	-3214.474
D3	0.4	30.24059	6.683678	6.706421	6.652781	**10.19771**
D3	0.8	71.14199	20.2795	19.89706	19.81336	**11.15207**
D3	0.95	263.2076	83.65639	79.13363	79.24412	**21.17346**
D3	0.995	2396.589	721.5145	727.141	863.4967	**84.43648**

5. CONCLUSIONS

We argue that quantile allocations can be useful to identify graphically some unique characteristics of multivariate distributions: to some extent, the trace of λ's or f's is a sort of fingerprint of each distribution, which entails characteristics of both the dependence structure and the shape and scale of the marginals. Plotting λ or f provide information about super- or sub-additivity of the VaR when aggregating the marginals. This issue is very relevant for risk managers in practice. An additional advantage of these kind of analysis is that compiles information of multi-dimensional randomness (random vectors) into only two-dimensions (α and λ). There is room for further research on this topic. For instance, identifying mathematically the partial contribution of the dependence structure in λ's.

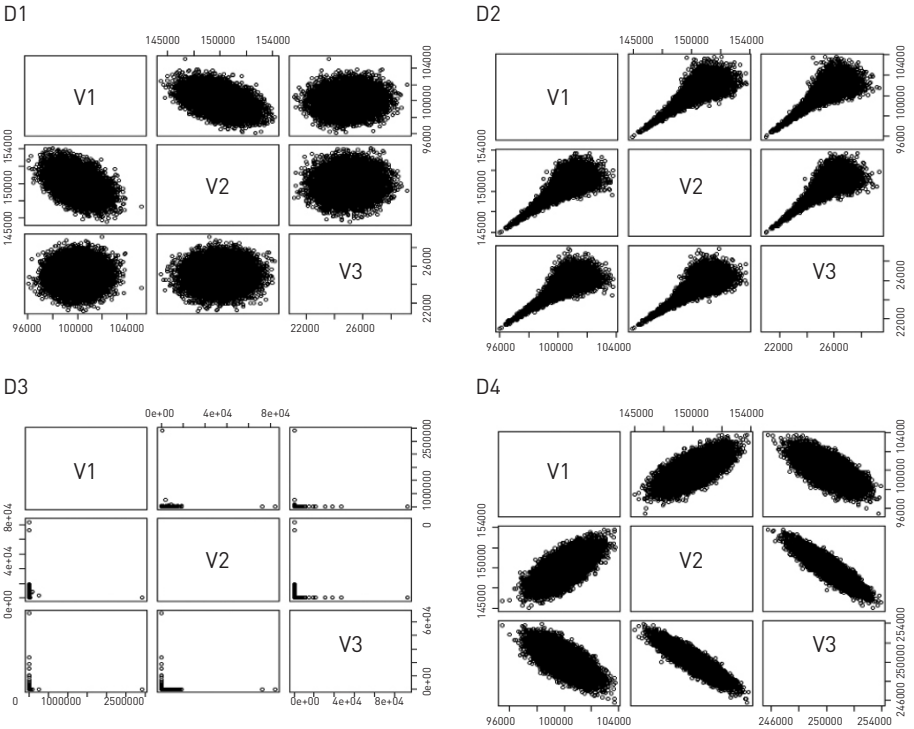
ACKNOWLEDGMENTS

The Spanish Ministry of Science and Innovation supported this study under grant PID2019- 105986GB-C21, as did the Secretaria d'Universitats i Recerca of the Catalan Government under grant 2020-PANDE-00074.

ANNEX

Scatterplots of the four multivariate distributions considered

Figure A.1. Scatterplots of the four multivariate distributions



REFERENCES

Belles-Sampera, J., Guillen, M., and Santolino, M. (2022). "Risk capital allocation as an optimization problem". IREA Working paper, 1, 1,1-18.

Cai, J., and Wang, Y. (2021). "Optimal capital allocation principles considering capital shortfall and surplus risks in a hierarchical corporate structure". Insurance: Mathematics and Economics, 100, 329-349.

Denault, M. (2001). "Coherent allocation of risk capital". Journal of Risk, 4, 1, 1-34.

Dhaene, J., Goovaerts, M.J., and Kaas, R. (2003). "Economic capital allocation derived from risk measures". North American Actuarial Journal, 7, 2, 44-56.

Dhaene, J., Tsanakas, A., Valdez, E.A., and Vanduffel, S. (2012). "Optimal capital allocation principles". Journal of Risk and Insurance, 79, 1, 1-28.

Furman, E., and Zitikis, R. (2008). "Weighted risk capital allocations". Insurance: Mathematics and Economics, 43, 2, 263-269.

McNeil, A.J., Frey, R., and Embrechts, P. (2005). Quantitative Risk Management. Princeton Series in Finance. Princeton University Press, New York.

Tsanakas, A. (2009). "To split or not to split: Capital allocation with convex risk measures". Insurance: Mathematics and Economics, 44, 2, 268-277.

van Gulick, G., De Waegenaere, A., and Norde, H. (2012). "Excess based allocation of risk capital". Insurance: Mathematics and Economics, 50, 1, 26-42.

Zaks, Y., Frostig, E., and Levikson, B. (2006). "Optimal pricing of a heterogeneous portfolio for a given risk level". ASTIN Bulletin, 36, 1, 161––85.

Zaks, Y., and Tsanakas, A. (2014). "Optimal capital allocation in a hierarchical corporate structure". Insurance: Mathematics and Economics, 56, 48-55.

ORDEN EN VALORES EN RIESGO DE COLA SOBRE p_0

Alfonso José Bello
Universidad de Cádiz, Departamento de Estadística e I.O., España
alfonsojose.bello@uca.es

Julio Mulero
Universidad de Alicante, Departamento de Matemáticas, España
julio.mulero@ua.es

Miguel Angel Sordo
Universidad de Cádiz, Departamento de Estadística e I.O., España
mangel.sordo@uca.es

Alfonso Suárez-Llorens
Universidad de Cádiz, Departamento de Estadística e I.O., España
alfonso.suarez@uca.es

RESUMEN

El valor en riesgo de cola a un nivel p con p, es una medida de riesgo que captura el riesgo de cola de las distribuciones de pérdidas y retornos de activos más allá del cuantil p. Dadas dos distribuciones, el valor en riesgo de cola puede usarse para decidir cuál de las anteriores distribuciones es la más arriesgada. Cuando los valores en riesgo de cola de dos distribuciones quedan ordenados en el mismo sentido para todos los niveles de probabilidad p del intervalo (0,1), indica que una de las distribuciones es más arriesgada que la otra. Sin embargo, al precisar una condición tan exigente, posibilita que no podamos comparar dos distribuciones, aunque nuestra intuición indique que una de ellas es menos arriesgada que la otra. En este trabajo, estudiaremos una familia de órdenes estocásticos indexados por un nivel p_0, con $p_0 \in (0,1)$, que solo requiere el orden de los valores en riesgo de cola para todos los

niveles p por encima de p_0. Presentaremos sus principales propiedades y la compararemos con otras familias de órdenes estocásticos de la literatura. Por último, ilustraremos los resultados mediante un ejemplo con datos reales.

1. MOTIVACIÓN Y PRELIMINARES

En finanzas y ciencias actuariales, es habitual que los inversores y gestores de riesgos se preocupen por el riesgo de la cola derecha de las distribuciones (ver Wang [1]). Un método para comparar dos riesgos, X e Y, es el uso de ordenaciones estocásticas, pudiéndose concluir que X es menor que Y cuando una familia de medidas de riesgo (y no una sola medida) está en acuerdo con la conclusión de que X es menos arriesgada que Y. Por ejemplo, el orden creciente convexo requiere del orden de los valores en riesgo de cola para cualquier nivel $p_0 \in (0,1)$. Sin embargo, este enfoque tiene el inconveniente de que algunas distribuciones no se pueden comparar, a pesar de que intuyamos que una es menos arriesgada que la otra.

Una forma de aumentar el número de distribuciones que se pueden comparar es reduciendo el rango de valores del nivel p requeridos para estar en acuerdo con el orden. Por ejemplo, un inversor preocupado en los riesgos de la cola derecha, puede pensar que X es menos arriesgado que Y si los respectivos valores en riesgo de cola están ordenados en acuerdo con la conclusión de que X es menos arriesgado que Y para cualquier nivel p, tal que $p > p_0$, donde $p_0 \in (0,1)$ es elegido en función de sus preferencias.

Sea X una variable aleatoria de pérdidas de un activo financiero y sea F su función de distribución (pérdidas negativas son ganancias). El valor en riesgo de X a un nivel $p_0 \in (0,1)$, o p-cuantil, se define como

$$VaR_p[X] = F^{-1}(p) = inf\{x: F(x) \geq p\}, \text{ para todo } p \in (0,1).$$

Para un $p_0 \in (0,1)$, $VaR_p[X]$ representa la máxima pérdida que el inversor puede sufrir con una confianza del $100p\%$. Sin embargo, el VaR no describe el comportamiento de la cola más allá del nivel y, además, no es subaditivo. Una alternativa al VaR, sin dichas limitaciones, es el valor en riesgo de cola o TVaR, definido por

$$\boldsymbol{TVaR_p[X] = \frac{1}{1-p}\int_p^1 F^{-1}(u)\,du}, \text{ para todo } p_0 \in (0,1).$$

Si X es continua, $TVaR_p[X] = E[X|X > F^{-1}(p)]$. $TVaR_p[X]$ representa el promedio de las pérdidas cuando estas superan el $VaR_p[X]$.

En este trabajo, estudiamos la siguiente familia de órdenes estocásticos indexados por niveles $p_0 \in (0,1)$, la cual fue presentada con detalle por Bello et al. (2020).

Definición 1.1 (Bello et al., 2020) Sean X e Y dos variables aleatorias y sea $p_0 \in (0,1)$. Entonces, se dice que X es menor que Y en el orden p_0-valor en riesgo de cola, denotado como $X \geq_{p0-tvar} Y$, si

$$TVaR_p[X] \leq TVaR_p[Y], \text{ para todo } p > p_0.$$

Cuando la propiedad se cumple para $p_0 = 0$, se tiene el orden creciente convexo usual (Lema 2.1 en Sordo y Ramos, 2017), denotado por $X \geq_{icx} Y$. Los libros de Shaked y Shanthikumar (2007) y Belzunce et al. (2016) estudian con detalle el orden creciente convexo. Si $X \geq_{icx} Y$, entonces X es más pequeño y menos variable que Y. En finanzas y ciencias actuariales, el orden creciente convexo es habitualmente interpretado en términos de contratos/órdenes de limitaciones de pérdidas. Específicamente, se cumple que $X \geq_{icx} Y$ si, y solo si,

$$E[(X - x)_+] \leq E[(Y - x)_+], \text{ para todo } x \in R,$$

donde $(x)_+ = x$, **s**i $x \geq 0$ y $(x)_+ = 0$, si $x < 0$, o, equivalentemente, si

$$\int_x^{+\infty} \overline{F}(t)\,dt \leq \int_x^{+\infty} \overline{G}(t)\,dt, \text{ para todo } x \in R,$$

siempre que las integrales existan, donde $\overline{F} = 1 - F$ y $\overline{G} = 1 - G$ son las respectivas funciones de supervivencia de X e Y.

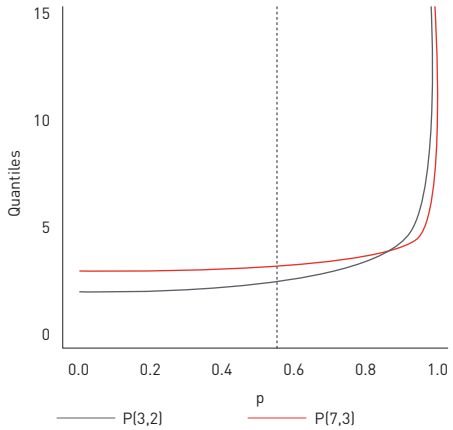
Veamos un ejemplo analítico que motiva el estudio de la nueva familia de órdenes estocásticos. Sean dos distribuciones de pérdidas Pareto, $X \sim P(7,3)$ e $Y \sim P(3,2)$. El valor en riesgo de una distribución Pareto $Z \sim P(a,k)$, con parámetro de forma $a < 0$ y parámetro de escala $k \in R$, viene dado por

$$F_Z^{-1}(p) = \frac{k}{(1-p)^{\frac{1}{a}}}, \text{ para todo } p \in (0,1),$$

y, si $a < 1$, $E[Z] = ak/(a - 1)$. La Figura 1 muestra la gráfica de las funciones cuantiles de X e Y. Dado que $E[X] = 3.5 > 3 = E[Y]$, se tiene que $X \not\leq_{icx} Y$.

Sin embargo, se verifica que $TVaR_p[X] \leq TVaR_p[Y]$, para todo $p > 0.55482$ donde $p_0 > 0.55482$ es el menor valor tal que $X \leq_{p0-tvar} Y$. Un inversor preocupado por las grandes desviaciones ocasionadas por las pérdidas de la cola derecha evaluará a X como menos arriesgada que Y, aunque $E[X] > E[Y]$ y $X \not\leq_{icx} Y$.

Figura 1. Funciones cuantiles para $X \sim P(7,3)$ e $Y \sim P(3,2)$.



La idea de limitar el número de comparaciones no es nueva. Por ejemplo, Cheung y Vanduffel (2013) siguieron la misma idea, aunque considerando variables aleatorias con idéntica media. Para otros estudios sobre órdenes basados en comparaciones de colas, consultar Sordo et al. (2015), Mulero et al. (2017) y Belzunce et al. (2020).

En la Sección 2, estudiaremos algunas propiedades del orden $\leq_{p0-tvar}$ y daremos condiciones bajo las que se cumple el orden. También estudiaremos su relación con otros órdenes estocásticos conocidos y compararemos familias paramétricas de soluciones. En la Sección 3, aplicaremos la nueva familia de órdenes en un conjunto de datos reales. La Sección 4 contiene las conclusiones del trabajo.

En este trabajo, "creciente" es "no-decreciente" y "decreciente" es "no-creciente". Las variables aleatorias tienen media finita. Dada una función h, S^- (h) denota el número de cambios de signo de h en su soporte (soporte vacío descartado).

2. PROPIEDADES Y RELACIONES CON OTROS ÓRDENES ESTOCÁSTICOS

Claramente, $X \leq_{p0-tvar} X$ para todo $p_0 \in (0,1)$. También, $X \leq_{p0-tvar} Y$, para $p_0 \in (0,1)$, implica que $X \leq_{p0-tvar} Y$, para todo $q_0 \in [p_0,1)$. Otra propiedad es que si X, Y y Z son tres variables aleatorias tales que $X \leq_{p0-tvar} Y$ e $Y \leq_{p1-tvar} Z$, con $p_0, p_1 \in (0,1)$, entonces $X \leq_{p3-tvar} Z$, con $p_3 = máx\{p_0,p_1\}$. Finalmente, si $X \leq_{p0-tvar} Y$ e $Y \leq_{p1-tvar} X$, con $p_0, p_1 \in (0,1)$, entonces $TVaR_p[X] = TVaR_p[Y]$, para todo $p_0 \in [p_3,1)$, donde p_3 se define como en la propiedad anterior.

A continuación, presentaremos las propiedades de clausura bajo convergencia en distribución y bajo transformaciones crecientes convexas. Dada una secuencia de variables aleatorias $\{X_n : n = 1,2, ...\}$ y una variable aleatoria X, con respectivas funciones de distribución F_n y F, se dice que X_n converge en distribución a X, o $\boldsymbol{X_n \xrightarrow{d} X}$, si $\boldsymbol{\lim_{n\to\infty} F_n(x) = F}$, para todo x en el que F es continua.

Proposición 2.1 (Bello et al. 2020) Sean $\{X_n : n = 1,2, ...\}$ e $\{Y_n : n = 1,2, ...\}$ dos secuencias de variables aleatorias continuas y positivas tales que $\boldsymbol{X_n \xrightarrow{d} X}$ y $\boldsymbol{Y_n \xrightarrow{a} Y}$. X_n y X tienen soporte común para todo $n \in N$ y $\boldsymbol{\lim_{n\to\infty} E[X_n] = E[X]}$ (y, análogamente, para Y_n e Y). Si $X_n \leq_{p0-tvar} Y_n$ para $p_0 \in (0,1)$, entonces $X \leq_{p0-tvar} Y$.

Proposición 2.2 (Bello et al. 2020) Sean X e Y dos variables aleatorias y sea ϕ una función creciente y convexa. Si $X \leq_{p0-tvar} Y$, para $p_0 \in (0,1)$, entonces $\phi(X) \leq_{p0-tvar} \phi(Y)$.

A continuación, presentamos la relación entre el nuevo orden y el orden *icx* de variables aleatorias que representan una limitación de pérdidas en una inversión.

Proposición 2.3 (Bello et al. 2020) Sean X e Y dos variables aleatorias. Si, para $p_0 \in (0,1)$, máx$\{X, F^{-1}(p_0)\} \leq_{icx}$ máx$\{Y, G^{-1}(p_0)\}$, entonces $X \leq_{p0-tvar} Y$.

El siguiente resultado interpreta el orden p_0-valor en riesgo de cola en términos de contratos/órdenes de limitaciones de pérdidas.

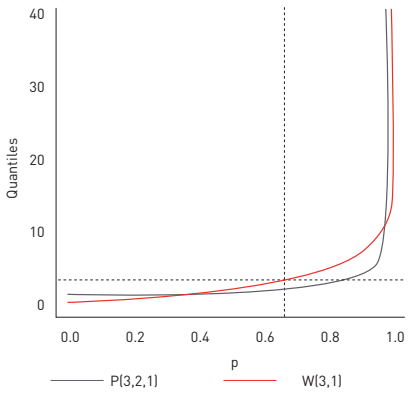
Proposición 2.4 (Bello et al. 2020) Sean X e Y dos variables aleatorias con respectivas funciones de distribución F y G.

(i) Si $X \leq_{p0-tvar} Y$, entonces $\int_x^{+\infty} \overline{F}(t)\,dt \leq \int_x^{+\infty} \overline{G}(t)\,dt$, para todo $x \geq F^{-1}(p_0)$.

(ii) Si $\int_x^{+\infty} \overline{F}(t)\,dt \leq \int_x^{+\infty} \overline{G}(t)\,dt$, para todo $x \geq x_0$, entonces $X \leq_{G(x0)-tvar} Y$.

Dadas dos variables aleatorias X e Y, ¿si $X \leq_{p0-tvar} Y$, para $p_0 \in (0,1)$, entonces $\int_x^{+\infty} \overline{F}(t)\,dt \leq \int_x^{+\infty} \overline{G}(t)\,dt$, para $x_0 \geq F^{-1}(p_0)$?. En general, la respuesta es no. Si $X \leq_{p0-tvar} Y$, para $p_0 \in (0,1)$, tal que $G^{-1}(p_0) < F^{-1}(p_0)$ y $\int_{p_0}^{1} F^{-1}(u)\,du = \int_{p_0}^{1} G^{-1}(u)\,du$, se puede obtener $x_0 \in (G^{-1}(p_0), F^{-1}(p_0))$ tal que $\int_{x_0}^{+\infty} (\overline{F}(t) - \overline{G}(t))\,dt > 0$. Similarmente, se demuestra que $\int_x^{+\infty} \overline{F}(t)\,dt \leq \int_x^{+\infty} \overline{G}(t)\,dt$, para todo $x \geq x_0$, no implica que $X \leq_{p0-tvar} Y$, para $p_0 < F(x_0)$. La Figura 2 ilustra dicha situación, para $X \sim W(3,1)$ (distribución Weibull) e $Y \sim P(3/2,1)$ (distribución Pareto), con respectivas funciones de distribución F y G. Recordar que si $X \sim W(\alpha,\beta)$ entonces $F^{-1}(p) = \alpha(-\log(1-p))^{1/\beta}$, para $p_0 \in (0,1)$. Se puede comprobar que $X \leq_{p0-tvar} Y$ para $p_0 = 0.68147$ y que $\int_x^{+\infty} \overline{F}(t)\,dt \leq \int_x^{+\infty} \overline{G}(t)\,dt$, para todo $x \geq F^{-1}(p_0) = 3.43711$. Sin embargo, como $G^{-1}(p_0) < F^{-1}(p_0)$ y $\int_{p_0}^{1} F^{-1}(u)\,du = \int_{p_0}^{1} G^{-1}(u)\,du$, no se cumple que $\int_x^{+\infty} \overline{F}(t)\,dt \leq \int_x^{+\infty} \overline{G}(t)\,dt$, para todo $x \geq x_0$, con $x_0 < F^{-1}(p_0)$.

Figura 2. Funciones cuantiles para $X \sim W(3,1)$ e $Y \sim P(3/2,1)$.



El siguiente corolario muestra la condición necesaria para obtener el "si, y solo si".

Corolario 2.5 (Bello et al. [2]) Sean X e Y dos variables aleatorias con respectivas funciones de distribución F y G, y sea $p_0 \in (0,1)$ tal que $F^{-1}(p_0) \leq G^{-1}(p_0)$. Entonces, $X \leq_{p0-tvar} Y$ si, y solo si, se cumple una de las siguientes condiciones equivalentes:

(i) $E[(X - x)_+] \leq E[(Y - x)_+]$, para todo $x \geq F^{-1}(p_0)$.

(ii) $\int_x^{+\infty} \overline{F}(t)\, dt \leq \int_x^{+\infty} \overline{G}(t)\, dt$, para todo $x \geq F^{-1}(p_0)$.

El Teorema 2.6.7 en Belzunce et al. (2016), junto con el siguiente teorema, permiten encontrar distribuciones paramétricas tales que $X \leq_{p0-tvar} Y$ para algún $p_0 \geq 0$ y $X \not\leq_{icx} Y$.

Teorema 2.6 (Bello et al. 2020) Sean X e Y dos variables aleatorias con respectivas funciones de distribución F y G. Si $S^-(G^{-1} - F^{-1})$ es finito, no nulo y el último cambio de signo ocurre de – a +, entonces $X \leq_{p0-tvar} Y$, donde p_n denota el último punto de cruce.

A continuación, se presentarán algunos ejemplos, ilustrados en la Figura 3.

Ejemplo 2.7 (Bello et al. 2020) En los siguientes ejemplos, $X \leq_{p0-tvar} Y$ y $X \not\leq_{icx} Y$.

(i) Sean $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$ dos variables aleatorias con distribuciones normales tales que $\mu_1 > \mu_2$ y $\sigma_1 > \sigma_2$. Entonces, $X \leq_{p0-tvar} Y$, donde $p_0 = F_Z\left(\frac{\mu_1-\mu_2}{\sigma_2-\sigma_1}\right)$ y $Z \sim N(0,1)$.

(ii) Sean $X \sim$ *Logística*(μ_1, σ_1) e $Y \sim$ *Logística*(μ_2, σ_2) dos variables aleatorias con distribuciones logísticas tales que $\mu_1 > \mu_2$ y $\sigma_1 > \sigma_2$. Entonces, $X \leq_{p0-tvar} Y$, donde $p_0 = F_Z\left(\frac{\mu_1-\mu_2}{\sigma_2-\sigma_1}\right)$ y $Z \sim$ *Logística* $(0,1)$.

(iii) Sean $X \sim W(\lambda_1, k_1)$ e $Y \sim W(\lambda_2, k_2)$ dos variables aleatorias con distribuciones Weibull tales que $E[X] > E[Y]$ y $k_2 < k_1$. Entonces, $X \leq_{p0-tvar} Y$, donde $p_0 = F_Z(a_0^{b0})$, $a_0 = \lambda_1/\lambda_2$, $b_0 = k_1 k_2/(k_1 - k_2)$ y $Z \sim W(1,1)$.

(iv) Sean $X \sim P(a_1,k_1)$ e $Y \sim P(a_2,k_2)$ dos variables aleatorias con distribuciones Pareto tales que $E[X] > E[Y]$ y $k_2 < k_1$. Entonces, $X \leq_{p0-tvar} Y$, donde $p_0 = F_Z(a_0^{b0})$, $a_0 = a_1/a_2$, $b_0 = k_1 k_2/(k_1 - k_2)$ y $Z \sim P(1,1)$.

De la Proposición 2.2, *(i) y (ii)* en el Ejemplo 2.7 también son válidos para las familias de distribución log-normal y log-logística, respectivamente.

Para concluir la sección, veremos que el nuevo orden es un orden puro de cola en el sentido de Rojo(1992). Es decir, si $X \leq_{p0-tvar} Y$, entonces la función de densidad de X decrece más rápido que la función de densidad de Y.

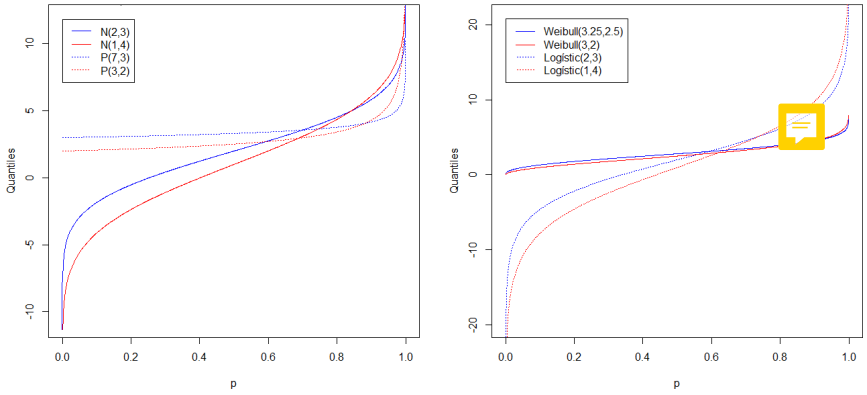
Proposición 2.8 (Bello et al. 2020) Sean X e Y dos variables aleatorias con respectivas funciones de densidad f y g. Sea $p_0 \in (0,1)$. Entonces,

$$X \leq_{p0-tvar} Y \text{ implica que } \lim_{x\to\infty} \frac{f(x)}{g(x)} \leq 1.$$

3. UN EJEMPLO CON DATOS REALES

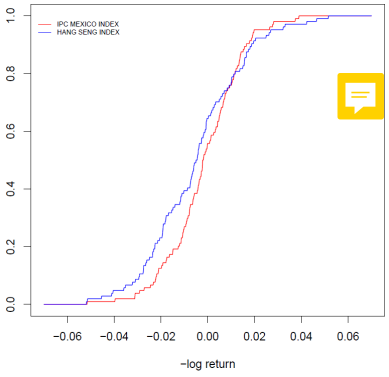
En esta sección, vamos a presentar una aplicación en finanzas con un conjunto de datos reales, que involucra dos variables aleatorias de –log-retornos. Se ha considerado, si p_t denota el precio de un activo el día t**,** el correspondiente –log-retorno del activo el día t, el cual queda definido mediante $r_t = -log\,(p_t/p_{t-1})$. Los datos son de acceso público y se pueden obtener de los repositorios de Yahoo! Finance. Para eliminar el efecto de dependencia en el tiempo, en los datos, se han registrado los cierres semanales de operaciones.

Figura 3. Funciones cuantiles. Izquierda: $X \sim N(2,3)$ e $Y \sim N(1,4)$ y $X \sim P(7,3)$ e $Y \sim P(3,2)$.Derecha: $X \sim$ *Logística*$(2,3)$ e $Y \sim$ *Logística*$(1,4)$ y $X \sim W(3.25,2.5)$ e $Y \sim W(3,2)$.



Hemos considerado dos índices bursátiles nacionales: el mexicano *S&P/BMV IPC* (Bolsa Mexicana de Valores), denotado por *MXX*, y el índice Hang Seng (Bolsa de Valores de Hong Kong), denotado por *HSI*. Para cada índice, se han obtenido muestras con los cierres semanales comprendidos entre el 1 de febrero de 2016 y el 31 de enero de 2018 ($n = 104$). Hemos denotado por R^{MXX} y R^{HSI} los –log-retornos de *MXX* y *HSI*, respectivamente. Vamos a obtener evidencias empíricas para concluir que $R^{MXX} \not\leq_{icx} R^{HSI}$, al mismo tiempo de que existe $p_0 \in (0,1)$ tal que $R^{MXX} <_{po-tvar} R^{HSI}$. Antes, hemos comprobado la aleatoriedad de los datos mediante un clásico test de rachas aplicado a R^{MXX} y R^{HSI} (p-valores 0.237 y 0.1149, respectivamente).

Figura 4. Funciones empíricas de las distribuciones de R^{MXX} y R^{HSI}.



La Figura 4 muestra evidencias empíricas de que $R^{MXX} <_{po-tvar} R^{HSI}$, para cierto nivel $p_0 \in (0,1)$ (por el Teorema 2.6). Por otro lado, vemos que $E[R^{MXX}] > E[R^{HSI}]$. Para ello, primero concluimos que ambas distribuciones, R^{MXX} y R^{HSI}, son simétricas (p-valores de contrastes *M*, *CM* y *MMG* del paquete *lawstat* de *R* mayores que 0.05). A continuación, concluimos que la mediana de R^{MXX} es mayor que la de R^{HSI} (p-valor de contraste de signos por rangos de Wilcoxon para muestras pareadas de 0.01). Por tanto, como $E[R^{MXX}] > E[R^{HSI}]$, $R^{MXX} \not\leq_{icx} R^{HSI}$.

Dos distribuciones clásicas para ajustar log-retornos son la distribución normal y la distribución logística (refleja mejor un exceso de curtosis). La Tabla 1 contiene los p-valores de los contrastes *K-S* para el ajuste de ambas distribuciones, y la Figura 5 muestra los histogramas de los log-retornos, junto a sus correspondientes densidades estimadas (parámetros estimados en la Tabla 2). Aunque la Tabla 1 indica que las distribuciones normales se ajustan mejor, ambas opciones son apropiadas. Por el Ejemplo 2.7, existe suficiente evidencia para suponer que $R^{MXX} <_{po-tvar} R^{HSI}$, y el punto de cruce podría ser calculado. Como conclusión, un inversor preocupado en los valores en riesgo de cola a niveles superiores que p, evaluará R^{MXX} como menos arriesgada que R^{HSI}, aunque $E[R^{MXX}] > E[R^{HSI}]$ y $R^{MXX} \not\leq_{icx} R^{HSI}$.

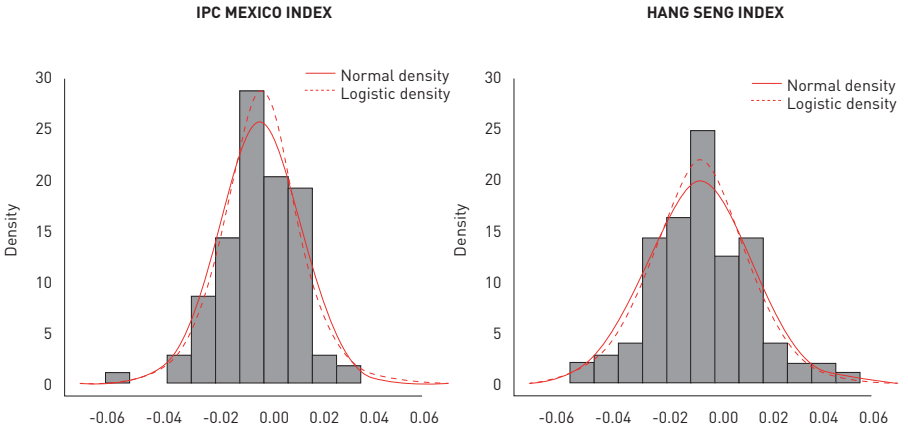
Tabla 1. p-valores resultantes de ajustar R^{MXX} y R^{HSI} a una distribución normal y logística.

K-S Goodness of Fit Test (p-Value)		
Index	**Normal**	**Logistic**
MXX	0.0512	0.0347
HSI	0.0525	0.0557

Tabla 2. Estimaciones de máxima verosimilitud (*MLE*) para distribuciones normales y logísticas.

	Normal		**Logistic**	
Index	μ	σ	μ	σ
MXX	−0.001486327	0.01556418	−0.001226663	0.008706823
HSI	−0.005058756	0.02017290	−0.005402021	0.011434130

Figura 5. Histogramas de los log-retornos con su correspondiente densidad estimada normal (línea continua) y su correspondiente densidad estimada logística (línea punteada), superpuestas. La gráfica de la izquierda corresponde a R^{MXX} y la derecha a R^{HSI}.



5. CONCLUSIONES

En este trabajo, se ha introducido una familia de órdenes estocásticos, indexada a niveles $p_0 \in (0,1)$, que resulta útil cuando estamos interesados en los riesgos de la cola derecha. Fijado p_0, decimos que X es menos arriesgado que Y si el valor en riesgo de cola de X es menor que el valor en riesgo de cola de Y para cualquier nivel p tal que $p > p_0$. Hemos estudiado las propiedades de esta familia de órdenes así como sus relaciones con otros órdenes estocásticos. Hemos ilustrado los resultados con un conjunto de datos financieros reales de log-retornos.

AGRADECIMIENTOS

Los autores agradecen la financiación del Ministerio de Ciencia e Innovación (proyecto PID2020-116216GB-I00).

BIBLIOGRAFÍA

Bello, A.J., Mulero, J., Sordo, M.A., and Suárez-Llorens, A. (2020). "On partial stochastic comparisons based on Tail Values at Risk". Mathematics, 8, 1181.

Belzunce, F., Franco-Pereira, A.M., and Mulero, J. (2020). "New stochastic comparisons based on tail value at risk measures". Communications in Statistics - Theory Method, doi:10.1080/03610926.2020.1754857.

Belzunce, F., Martínez-Riquelme, C., and Mulero, J. (2016). An Introduction to Stochastic Orders. Academic Press, Elsevier Ltd.: London, UK.

Cheung, K.C., and Lo, A. (2013). "Characterizations of counter-monotonicity and upper comonotonicity by (tail) convex order". Insurance, Mathematics and Economics, 53, 334-342.

Cheung, K.C., and Vanduffel, S. (2013). "Bounds for sums of random variables when the marginal distributions and the variance of the sum are given". Scandinavian Actuarial Journal, 2, 103-118.

Mulero, J., Sordo, M.A., de Souza, M.C., Suárez-Llorens, A. (2017). "Two stochastic dominance criteria based on tail comparisons". Applied Stochastic Models in Business and Industry, 33, 6, 575-589.

Rojo, J. (1992). "A pure tail ordering based on the ratio of the quantile functions". Annals of Statistics, 20, 570-579.

Shaked, M., and Shanthikumar, J.G. (2007). Stochastic Orders. Springer Series in Statistics. Springer: New York, NY, US.

Sordo, M.A., and Ramos, H.M. (2007). "Characterization of stochastic orderings by L-functionals". Statistical Papers 2007, 48, 249-263.

Sordo, M.A., de Souza, M.C., and Suárez-Llorens, A. (2015). "A new variability order based on tail-heaviness". Statistics, 49, 1042-1061.

Wang, S. (1998). "An actuarial index of the right-tail risk". North American Actuarial Journal, 2, 88-101.

MODELLING UNOBSERVED HETEROGENEITY BASED ON FINITE MIXTURE OF REGRESSIONS AND A CLASSIFICATION RULE

Lluís Bermúdez
Universitat de Barcelona, Riskcenter-IREA, Spain
lbermudez@ub.edu

Dimitris Karlis
Athens University of Economics and Business, Greece
karlis@aueb.gr

ABSTRACT

A natural way to deal with unobserved heterogeneity in ratemaking is to consider *k*-finite mixtures of some typical regression models. This approach allows for overdispersion and the zero-inflated model represents a special case and, simultaneously, allows for an interesting interpretation of the subgroups derived from the typical clustering interpretation of the finite mixture models. However, current approaches are limited by the fact that the models are not fully usable for prediction since the observed number of claim counts is not known a priori and, hence, it is rather impossible to classify new clients into subgroups. In this paper, we tackle this problem using a classification rule to classify new clients to one of the *k* subgroups of risks. We discuss practical issues about this with a real data set. The main finding is that the *k* subgroups of risks defined by the finite mixture exhibit different regression structures, which provides much better fit and predictive accuracy.

1. INTRODUCTION

Pricing automobile insurance is specially complicated because of the presence of very heterogeneous portfolios. One way to handle this feature consists of

segmenting the portfolio in homogenous classes so that all the insured who belong to a class pay the same premium (Denuit et al., 2007).

To do this, a ratemaking based on generalized linear models (GLM) is usually accepted. The first common use of a GLM for modelling the frequency component of the premium is the Poisson regression model and its generalizations. However, in automobile insurance datasets, when modelling claims counts for ratemaking purposes, there is a problem with the unobserved heterogeneity.

The unobserved heterogeneity is caused by differences in driving habits among policyholders that cannot be observed or measured and, hence, included in the premium by the actuary (for example, driving ability, driving aggressiveness or the degree of obeying traffic regulations).

This often leads to overdispersion and a relatively large number of zeros which cannot be fully remedied by Poisson regression models. Many attempts have been made in the actuarial literature to account for this unobserved heterogeneity. For example: zero-inflated, hurdle, and compound frequency models.

However, a natural way to deal with unobserved heterogeneity is to consider mixtures of a simpler models. In this paper, we consider k-finite mixtures of some typical regression count models (finite mixtures of regression or, in short, FMR models), see Park and Lord (2009). We may assume a finite mixture of Poisson (or negative binomial) regressions. In Bermudez et al. (2020), this approach is applied to a car insurance claims dataset, obtaining the following conclusions. First, finite mixture of regressions models (Poisson or negative binomial) produced better fit than classical models like compound and zero-inflated models. Second, the portfolio was comprised of two groups of policyholders or drivers that we can call "good" drivers and "bad" drivers according to their driving habits or behaviour. Third what is even more important for ratemaking purposes, the two groups of policyholders exhibited different regression structures, in other terms, they behave in different ways with regard to the a priori factors. And finally, the two groups of policyholders presented very different expected claim frequency values.

However, a problem arises when we want to apply these models for predicting a new client. Since Y (the number of claims) is not available (only x's or a priori factors are available) we cannot use these models to assign a new observation to a group and, hence, take advantage for "better" identifying the group (and the characteristics) that the new client belongs. This invalidates the advantage of this approach since we still predict by the overall mean and not using the corresponding regression structure of each group.

It is worth noting that is not an easy problem and a little is known about this in the literature. One possible approach could be to assign covariates on the mixing proportions, but then problems of identifiability may arise. Therefore, in this paper, we propose a classification-based approach.

2. A CLASSIFICATION-BASED ALGORITHM

At this point, we are ready to propose a generic solution. First, we fit the FMR model using the available data (we call the available data as training data). Second, for each new observation (we call them as test data), we find observations from the training data that are closer (based on some distance) to the new observation, looking only the x's and not the response. We denote this set of observations as S_k. Third, from the fitted FMR, we calculate the posterior probabilities for all observations in this set S_k and find which of the groups is more probable. Finally, we assign the new observation to that group.

Of course, we can consider different variants of this generic algorithm. For example, by defining different distances or by using different assignment methods.

Following the application section in Bermudez et al. (2020), the available data are such that the response variable represents the total number of claims reported by policyholders, and the seven covariates (or a priori factors) describe some of their characteristics, commonly used in ratemaking. Note that in this case, all covariates are binaries and, hence, we can have a finite number of combinations and use zero distance criterion.

From the generic proposal discussed before, we can adapt the algorithm to our data:

- Step 1: Fit a k-finite mixture of regression model to the training data and assign each observation to any of the k components using the posterior probabilities. Note that for the training data we have observed the response Y, so we can calculate the posterior probabilities.

- Step 2: Define a similarity index between the new observation and the ones in the training data, this depends solely on the structure of their covariate information. Since we have only binaries covariates, a matching coefficient can be used. So, only training data with the exact values for the 7 covariates are considered for this (zero distance criterion).

- Step 3: For each observation of the training data with the same covariate's values, see what classification results based on the posterior probabilities.

- Step 4: Assign the new observation to the component that most of the observations with the same covariate information in the training data have been assigned.

- Step 5: Predict the outcome of the new observation using the GLM model of the component where the new observation has been assigned at Step 4.

3. DATA AND RESULTS

In our case, the entire dataset, of more than eighty thousand observations, has been randomly separated into a training dataset (75% of the full data) and a test dataset (25% of the full data). The training dataset is used for the parameter estimation of the models whereas the test dataset for measuring their predictive accuracy. As a prediction accuracy metric, the Standardized Mean Absolute Prediction Error (SMAPE) criterion was selected.

Table 1. SMAPE values for FMPR models using either algorithm predictions orweighted-average predictions.

Models	Algorithm Predictions	Weighted-Average Predictions
Poisson	-	1.7010
2-FMPR	1.2649	1.7005

When fitting a finite mixture of Poisson regression model (FMPR) to this data set, we can clearly see that if we use a weighted-average prediction, we will not differentiate the premiums for the two groups and we will place in some space in between, as in the simple Poisson regression model. On the contrary, with the classification-based approach and the algorithm here proposed, we will get a better separation by predicting using the corresponding score, and hence, obtaining a more adjusted ratemaking. This better performance can be seen in Table 1. When we predict using a weighted-average between 2 components, the 2-FMPR model produced almost the same value than the predictions obtained with the simple Poisson regression model. However, using the algorithm to classify new observations in one of the 2 groups and predicting with the corresponding score, SMAPE improved a lot.

4. CONCLUSIONS

In Bermúdez et al. (2020), we showed the advantages of using FMR models with respect to other models: 1) with respect to compound frequency models, FMR models account for unobserved heterogeneity more effectively, providing a better fit and a better picture for managerial matters; 2) with respect to zero-inflated and hurdle models, the problem of unobserved heterogeneity is addressed in a better way, trying to fix overdispersion and excess of zeros at the same time; and 3) with respect to both, FMR models have a richer regression structure, with one score per component, that allow us to account jointly for both, observed and unobserved factors.

Assuming that these advantages will be of interest to actuaries, we have developed an extension to enhance the applicability of FMR models. Using the classification-based algorithm proposed here, we can benefit from the multiple scores provided by FMR models and obtain a much better predictive accuracy.

This approach can be applied to certain other finite mixtures of regression settings, including bivariate or multivariate models for claim counts. In Bermúdez and Karlis (2012), a bivariate finite mixture regression model for modelling two types of claims simultaneously was proposed.

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Science, Innovation and Universities (grant PID2019-105986GB-C21). Likewise, we want to acknowledge the support received by AGAUR of the Catalan Government (Grant 2017SGR1147).

REFERENCES

Bermúdez, L., and Karlis, D. (2012). "A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking". Computational Statistics & Data Analysis, 56, 3988-3999.

Bermúdez, L., Karlis, D., and Morillo, I. (2020). "Modelling unobserved heterogeneity in claim counts using finite mixture models". Risks, 8, 1, 10.

Denuit, M., Marechal, X., Pitrebois, S., and Walhin, J.-F. (2007). Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems. New York: Wiley.

Park, B-J., and Lord, D. (2009). "Application of finite mixture models for vehicle crash data analysis". Accident Analysis and Prevention,41, 683-691.

MAXIMUM DEPENDENCE LIMITS BETWEEN FREQUENCY AND SEVERITY USING EXPONENTIAL KERNELS IN SARMANOV DISTRIBUTION

Albert Martos Ramírez
Departament d'Econometria, Estadística i Economia Aplicada, RISKcenter, Institut de Recerca en Economia Aplicada (IREA), Universitat de Barcelona (UB), Spain

Catalina Bolancé
Departament d'Econometria, Estadística i Economia Aplicada, RISKcenter, Institut de Recerca en Economia Aplicada (IREA), Universitat de Barcelona (UB), Spain
bolance@ub.edu

ABSTRACT

The multivariate Sarmanov is a flexible distribution which allows us to model the dependence structure between variables with different distributions, including continuous or discrete random variables. This dependency can take any value restricted to a limited interval, which its minimum and maximum can be found by using kernels and the parameters of the discrete and continuous random variables. The use of exponential kernels can help us to relax these limits. Each kernel depends on the Laplace transform of the marginal distribution and thus we study the values of the frequency parameter of this transform which maximise the dependence interval width. The study is carried out in the context of the collective risk model and assuming dependence between claim frequency and severity. We assume that severity is Gamma distributed and frequency is Negative Binomial distributed, although the same study can easily be generalized for other distributions.

1. INTRODUCTION

Let S denote the aggregate claims, such as:

$$S = NX \tag{1}$$

where N is the number of claims and X the corresponding average claim size of a portfolio or a certain policy. We assume if $N = 0$ then $X = 0$.

In the classical collective model context to calculate insurance risk premium, claim frequency and severity independence can be assumed, which provides greater ease to the corresponding calculations. In practice, however, claim frequency and severity tend to be correlated, albeit this is usually small and positive or negative sign can be justified. We propose to use Sarmanov distribution with exponential kernels to model this dependence (Vernic et al., 2022; Alemany et al., 2021; Bolancé and Vernic, 2020). The main criticism of this distribution is that it imposes limits on dependency. We analyse how wide the limits of this dependence can be, depending on the frequency parameters of the Laplace transforms of the marginal distributions.

We study the particular case of the bivariate Sarmanov distribution with exponential kernels and considering Negative Binomial (NB) and Gamma (Ga) marginal distributions, hereafter bivariate Sarmanov-NB-Gamma distribution. These marginal distributions were used in other works as Vernic et al. (2022), Garrido et al. (2016) and Czado et al. (2012). We present the results of a application using a sample from two real data sets of auto insurance portfolios.

The rest of the paper is organized as follows: in Section 2, we briefly describe the proposed Sarmanov distribution, its principal properties and particular cases. In Section 3, we present the results of dependence limits and illustrate them using some particular cases for Sarmanov-NB-Gamma distribution. An application using a real data set containing auto insurance number and average cost of claims is discussed in Section 4. Finally, we conclude in Section 5.

2. THE BIVARIATE SARMANOV DISTRIBUTION

The bivariate Sarmanov distribution allows us to join variables with different distributions. Its bivariate probability distribution function (pdf) is (see Vernic et al., 2022):

$$f_{X,N}(x,n) = \begin{cases} p(0), \ n = x = 0 \\ p(n)f(x)\big(1 + \omega\psi(n,\delta)\phi(x,\gamma)\big), \ n \geq 1, \ x > 0 \end{cases}. \tag{2}$$

In the above formula $p(\cdot)$ is the probability mass function (pmf) of the discrete variable N and $f(\cdot)$ is the pdf of the continuous variable X. Furthermore, in the Sarmanov distribution ω is the dependence parameter and $\psi(n,\delta) = e^{-\delta n} - \frac{\mathcal{L}_N(\delta)-p(0)}{1-p(0)}$ and $\phi(x, \gamma) = e^{-\gamma x} - L_X(\gamma)$ are the two exponential kernel functions, where L denotes the Laplace transform of the distribution, where $X = X|X > 0$. Note that, by construction, for the Negative Binomial we work with the truncated Laplace transform.

The main concern with Sarmanov distribution is that we need to impose limits on the dependence parameter to guarantee that the expression defined in (2) is a density. This involves that we will not be able to fit large negative or positive dependence.

Let $m_1 = \inf_{n\geq 1} \psi(n,\delta)$, $m_2 = \inf_{x>0} \phi(x,\gamma)$, $M_1 = \sup_{n\geq 1} \psi(n,\delta)$ and $M_2 = \sup_{x>0} \phi(x,\gamma)$, then the limits for the dependence parameter are:

$$\max\left\{-\frac{1}{m_1 m_2}, -\frac{1}{M_1 M_2}\right\} \leq \omega \leq \min\left\{-\frac{1}{m_1 M_2}, -\frac{1}{M_1 m_2}\right\}. \tag{3}$$

2.1. The bivariate sarmanov-nb-gamma distribution

Let $N \sim NB(r, p)$ and $X \sim Gamma(\alpha,\beta)$ be the marginal distributions in Sarmanov model. In this case we denote our analysed distribution as $(N,X) \sim Sarmanov - NB - Gamma(r, p, \alpha, \beta, \omega, \delta, \gamma)$. The exponential kernels for this distribution are:

$$\psi(n,\delta) = e^{-\delta n} - \frac{\mathcal{L}_N(\delta)-p(0)}{1-p(0)} = e^{-\delta n} - \frac{\left(\frac{p}{1-qe^{-\delta}}\right)^r - p^r}{1-p^r}, \tag{4}$$

$$\phi(x,\gamma) = e^{-\gamma x} - \mathcal{L}_{\bar{X}}(\gamma) = e^{-\gamma x} - \left(\frac{\beta}{\beta+\gamma}\right)^{\alpha}.$$

Using (4) we obtain: $m_1 = \psi(+\infty, \beta)$, $m_1 = \phi(+\infty, \beta)$, $M_1 = \psi(1, \delta)$ and $M_2 = \phi(1, \gamma)$.

3. LOWER AND UPPER LIMITS FOR DEPENDENCE

Note that working with exponential kernels adds two parameters to the distribution, this can be a troublesome in estimation process. A simple alternative that has been used in previous works (Bolancé and Vernic, 2020; Bolancé et al., 2020; Bolancé and Vernic, 2019) consists of fixing $\delta = \gamma = 1$. However, this could cause narrow intervals for the dependence parameter. Alternatively, we propose to estimate frequency Laplace transform parameters, taking into account that both parameters are limited above. We can calculate these upper limits for δ and γ from de expression (3) as follows:

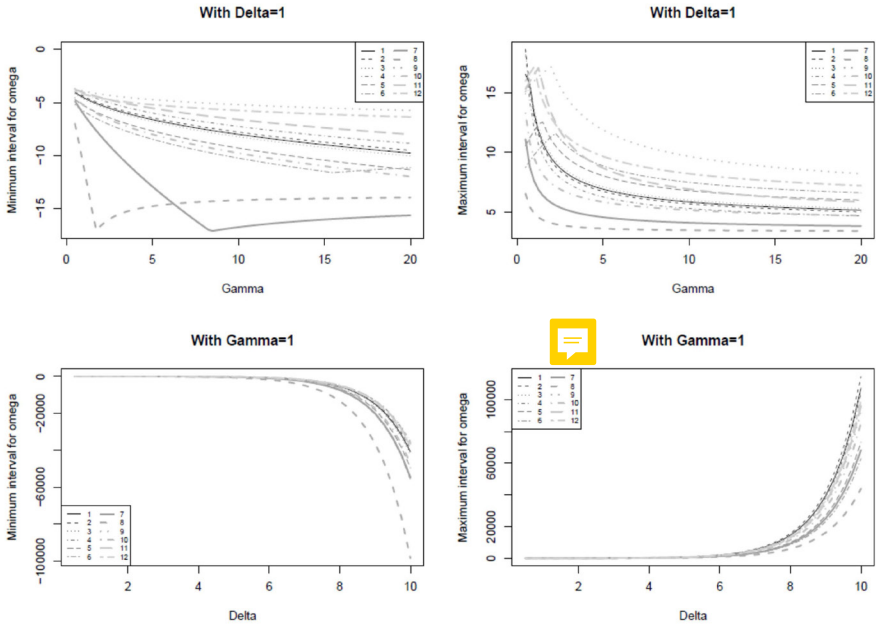
1. We minimise $\boldsymbol{D_1} = -\frac{1}{m_1 m_2}$ and $\boldsymbol{D_2} = -\frac{1}{M_1 M_2}$ and maximise $\boldsymbol{D_3} = -\frac{1}{m_1 M_2}$ and $\boldsymbol{D_4} = -\frac{1}{M_1 m_2}$ with respect to δ and γ. We denote $D_1^* = min_{\delta,\gamma} D_1$, $D_2^* = min_{\delta,\gamma} D_2$, $D_3^* = max_{\delta,\gamma} D_3$ and $D_4^* = max_{\delta,\gamma} D_4$.

2. Let $(\delta_{1*_\gamma}\ \gamma_{1*}) = \arg\min_{\delta,\gamma} D_1$ and $(\delta_{2*_\gamma}\gamma_{2*}) = \arg\min_{\delta,\gamma} D_2$ be the optimal values for the lower dependence limit, such as $(\delta_{*,}\gamma_*) = \{[\delta_{1*_\gamma}\ \gamma_{1*})]j = 1,2 : D_{1*} = \max(D_{1*},D_{2*})\}$, i.e. we select the maximum between the minimums.

3. Let $(\delta_{3,}^*\gamma_3^*) = \arg\max_{\delta,\gamma} D_3$ and $(\delta_{4,}^*\gamma_4^*) = \arg\max_{\delta,\gamma} D_4$ be the optimal values for the upper dependence limit, such as $(\delta^*_{,}\gamma^*) = \{[\delta_j^*\ \gamma_j^*)]j = 3,4 : D_j^* = \min(D_3^*,D_4^*)\}$, i.e. we select the minimum between the maximums.

4. The upper limits are $\delta_u = \max(\beta_{j*},\beta_j^*)$ and $\gamma_u = \max(\gamma_{j*},\gamma_j^*)$.

Figure 1. Lower and upper limits for dependence parameter in function of δ and γ.



In Figure 1 we plot the lower (on the left) and upper (on the right) limits of the dependence parameters for the 12 models that are analysed in Table 1. We show how for lower limits the larger δ and γ more negative is this bound. On the contrary, for upper limits δ and γ are inversely related.

In Table 1 we show some particular cases for some given values for the parameters of *Sarmanov – NB – Gamma*(r, p, α, β, ω, δ, γ). The upper limits for the frequency parameters are large. In practice we will not need such high frequency parameters, but these results allow us to analyse how these parameters could be increased. Increasing the frequency parameters of Laplace transforms also have a certain scale effect on the dependence parameter. Thus, the obtained limits for ω cannot be directly compared.

Finding a way to build a wider interval for the dependency parameter will turn the Sarmanov distribution easier to work with.

4. NUMERICAL EXAMPLE WITH REAL DATA

We estimate the Sarmanov-NB-Gamma distribution using a bivariate samples that contains claim frequency and severity for a random sample of 99,972 policyholders of a real auto insurance portfolio. In Table 2, we show the average severity according to the number of claims and the linear and rank correlation coefficients for $N, X > 0$ that indicates some positive dependence between frequency and severity.

Table 1. Limits for δ and γ. The pairs (δ_*, γ_*) minimise lower limits of dependence parameter and the pairs (δ^*, γ^*) maximise upper limits of dependence parameter.

Model ID.	(r,p,α,β)	δ_*	γ_*	δ^*	γ^*
1	(0.3,0.6,0.3,0.6)	10.0000	10.0000	12.9996	0.4997
2	(0.15,0.6,0.3,0.6)	10.0000	10.0000	12.9999	0.4999
3	(0.45,0.6,0.3,0.6)	10.0000	10.0000	10.0000	10.0000
4	(0.3,0.75,0.3,0.6)	10.0000	10.0000	12.9999	0.4999
5	(0.2,0.4,0.3,0.6)	11.0001	10.0009	10.0000	10.0000
6	(0.15,0.3,0.3,0.6)	11.0001	10.0011	10.0000	10.0000
7	(0.3,0.6,0.6,0.6)	11.0001	10.0037	14.0010	0.5008
8	(0.3,0.6,1.2,0.6)	15.4545	98.0551	13.9996	0.4996
9	(0.3,0.6,0.15,0.6)	10.0000	10.0000	10.0000	10.000
10	(0.3,0.6,0.3,0.3)	10.0000	10.0000	12.9998	0.4998
11	(0.3,0.6,0.3,1.2)	10.0000	10.0000	10.0000	10.0000
12	(0.3,0.6,0.15,0.3)	10.0000	10.0000	10.0000	10.0000

Table 2. Frequency and mean of the severity in function of number of claims.

Claims	Frequency	Mean Cost
0	92538	0
1	6166	543.0689
2	1122	2886.4754
3	125	2836.7892
4	18	2262.6000
5	3	1152.1233
Pearson	0.2800_*	
Kendall	0.2633_*	
Spearman	0.2946_*	

**p-value<0.0001 for statistical significance test*

In Table 3, we show the estimation results, the procedure is the same described in Vernic et al. (2022) but using three conditioned likelihood functions instead of only two. The first in function of the frequency parameter, given dependence and marginals parameters, the second in function of the dependence parameter given frequency and marginal parameters and the third in function of the marginal parameters given frequency and dependence parameters. The results indicate an improvement of fit when δ and γ are estimated with the rest of parameters, the AIC for free δ and γ is lower than the AIC for $\delta = \gamma = 1$.

Table 3. Estimated parameters for Sarmanov-NB-Gamma models.

	Free (δ, γ)	$\delta = \gamma = 1$
r	0.2897	0.2897
p	0.7655	0.7655
α	0.2741	0.2741
β	0.3962	0.3962
ω	2.4585	2.1470
δ	0.7011	1.0000
γ	0.6917	1.0000
Log-lik	-27338.342	-27348.769
AIC	54690.684	54707.538

5. CONCLUSIONS

In this paper, we have analysed as exponential kernel can help to flexibility the dependence limits of Sarmanov distribution, in the context of the collective risk model where we can assume dependence between frequency and severity. The same idea can be applied to alternative analysis that use different marginal distributions. Furthermore, we have studied as the frequency parameters of the Laplace transforms of the marginal distributions allow this interval to be extended. Select optimal frequency parameters improves clearly the fit of the Sarmanov-NB-Gamma model.

REFERENCES

Alemany, R., Bolancé, C., Rodrigo, R., and Vernic, R. (2021). "Bivariate mixed poisson and normal generalised linear models with Sarmanov dependence - an application to model claim frequency and optimal transformed average severity". Mathematics, 9. doi:10.3390/math9010073.

Bolancé, C., Guillen, M., and Pitarque, A. (2020). "A Sarmanov distribution with beta marginals: An application to motor insurance pricing". Mathematics, 8. doi:10.3390/math8112020.

Bolancé, C., and Vernic, R. (2019). "Multivariate count data generalized linear models: Three approaches based on the Sarmanov distribution". Insurance: Mathematics and Economics, 85, 89-103.

Bolancé, C., and Vernic, R. (2020). "Frequency and severity dependence in the collective risk model: An approach based on Sarmanov distribution". Mathematics, 8. doi:10.3390/math8091400.

Czado, C., Kastenmeier, R., Brechmann, E.C., and Min, A. (2012). "A mixed copula model for insurance claims and claim sizes". Scandinavian Actuarial Journal, 4, 278-305.

Garrido, J., Genest, C., and Schulz, J. (2016). "Generalized linear models for dependent frequency and severity of insurance claims". Insurance: Mathematics and Economics, 70, 205-215.

Vernic, R., Bolancé, C., and Alemany, R. (2022). "Sarmanov distribution for modeling dependence between the frequency and the average severity of insurance claims". Insurance: Mathematics and Economics, 102, 111-125.

INTRODUCING A NON-LINEAR REGRESSION MODEL IN AN EXCESS-OF-LOSS REINSURANCE CONTEXT

Emilio Gómez-Déniz
ULPGC, Departamento de Métodos cuantitativos en Economía y Instituto TiDes, España
Emilio.gomez-deniz@ulpgc.esl1,

Enrique Calderín-Ojeda
University of Melbourne, Departament of Economics, Australia
enrique.calderin@unimelb.edu.aul

ABSTRACT

This work discusses excess-of-loss reinsurance, paying particular attention to the reinsurance's position point of view. Under an excess-of-loss reinsurance arrangement, a claim is shared between the insurer and the reinsurer only if the claim exceeds a fixed amount, usually known as the excess or the retention level. We present a parametric model to calculate the excess-of-loss distribution, examining how covariates may be incorporated into the model and how these covariates may affect the expected payment. We will consider continuous and indicator covariates included in a real motor vehicle insurance portfolio. Our findings allow us to conclude that the decision of the retention level has an essential impact on some of the considered covariates. We believe this could help the reinsurance company improve its pricing policy. Finally, since claim size often increases over time due to inflation, it is worth investigating how this rate the reinsurer's payments.

1. INTRODUCCIÓN

The preface of Albrecher et al. (2017) states that *reinsurance is a fascinating field*. The reason for that is because reinsurance is a useful technique to solve various financial problems for insurance companies. As it is well-known by the actuarial community, a reinsurance contract establishes an agreement or contract between the reinsured and the reinsurer. In this contract, the latter compensates the former (also called the ceding company) a part of the risk subscribed by the reinsurer, with a third party (usually the customer) paying in return a reinsurance premium. This is the basic idea behind the type of reinsurance called excess-of-loss reinsurance. For a detailed study of reinsurance, in all its forms, it can be seen the works of Helselager (1993), Mata (2000) and Albrecher et al. (2017) and the references therein.

We present a parametric model to calculate the excess-of-loss distribution, examining how covariates may be introduced into the model and how these covariates may affect the expected payment. The obtained results allow us to conclude that the decision of the retention level has an essential impact on some of the considered covariates. We believe this could help the reinsurance company improve its pricing policy. Finally, since claim size often increases over time due to inflation, it is worth investigating how this affects the reinsurer´s payments.

The contents of this work are organized as follows. To do the work self-contained, we present in Section 2 the necessary mathematical tools used in the excess-of-loss reinsurance. Here, the effect of inflation is also briefly shown. A specific model with a distributional assumption for the loss is developed in Section 3. Finally, empirical applications are provided in Section 4. Additionally, some technical aspects are included in the Appendix of this work.

2. BACKGROUND

Let X the random variable denoting the amount of a claim. A reinsurance arrangement is an agreement between an insurer and a reinsurer under which claims that occur in a fixed period of time, for example one year, are split between the insurer

and the reinsurer in an agreed manner. Thus, the insurer is effectively insuring part of a risk with a reinsurer and, of course, pays a premium to the reinsurer for this cover. One effect of reinsurance is that it reduces the variability of claim payments by the insurer. That is, $X=Y+Z$, where X and Z are the random variables denoting the amount paid by the insurer and reinsurer, respectively.

Under an excess-of-loss reinsurance model the loss is shared between the insurer and the reinsurer if the amount of the expense exceeds a fixed quantity, say $q>0$. Otherwise, the insurer assumes the loss. Mathematically, this situation can be represented as $Y = min\{X, q\}$, $Z = max\{0, X-q\}$ and obviously we have that $X=Y+Z$. Observe that the loss are shared again between the insurer and the reinsurer but this sharing can have different proportions between origin and destination.

Moments of Z can be computed by using

$$E(Z^n) = \int_0^\infty (\max\{0, x-q\})^n f_X(x)dx = \int_q^\infty (x-q)^n f_X(x)dx$$

and in particular the expected value of the ceded amount is obtained by assuming $n=1$.

Let now Z denoting the amount of non-zero payment assumed by the reinsurer. The pdf of this random variable is equal to that of Z|Z>0 and it is given (see Dickson, 2005) $f_Z(z)= f_X(z + q)/\overline{F}_X(q)$, where $\overline{F}_X(\,)$ represents the survival function.

Prices often increase over time due to inflation, and it is worth investigating how this affects typical payments of the retained portion of claims by the insurer and the ceded portion to the reinsurer. For the proportional model, the matter seems to be simple but for the non-proportional model, some decisions that produced payments below the quantity q will now lead to additional payments. Furthermore, the effect of inflation produces some effect on the reinsurer's payment if we keep the threshold q fixed.

Suppose now that $\tau \geq 1$ is the inflation. Now, we have that

$$Y^* = \begin{cases} X, & x \leq q, \\ q, & X > q. \end{cases} \quad \text{and} \quad Z^* = \begin{cases} 0, & x \leq q, \\ \tau X - q, & X > q. \end{cases}$$

It is simple to see that the pdf of Z^* which results

$$f_{Z^*}(z^*) = \frac{f_X((z^* + q)/\tau)}{\tau \overline{F}(q/\tau)}.^*$$

Note that $\tau = 1$, is the case where non-inflation is assumed.

3. SPECIFIC MODEL

A gamma distribution with shape parameter 2 and scale parameter $\lambda > 0$ and with pdf given by $f_X(x) = \lambda_2 x \exp(\lambda x)$, $x > 0$, $\lambda > 0$, was assumed for the random variable X.

On the other hand, assuming inflation we have that, by using (1), we have that

$$f_Z(z) = \frac{\lambda^2 (z+q)\exp(-\lambda z/\tau)}{\tau(\tau + \lambda q)}^*, \quad z > 0,$$

with mean given by

$$E(Z) = \frac{\tau(2\tau + \lambda q)}{\lambda(\tau + \lambda q)}.$$

Observe that, from (3) we have that

$$\frac{d}{dq}E(Z) < 0, \quad \frac{d}{d\lambda}E(Z) < 0 \quad \text{and} \quad \frac{d}{d\tau}E(Z) > 0.$$

For purposes of including covariates it is convenient to take

$q = \tau(\Theta\lambda - 2\tau)/[\lambda(\tau - \Theta\lambda)]$. Now, we have that (2) can be rewritten as

$$f_Z(z) = \frac{\lambda}{\tau^3}\left[2\tau^2 + \theta\lambda^2 z - \lambda\tau(\theta + x)\right]\exp\left(-\frac{\lambda z}{\tau}\right), \quad z > 0.$$

In order to ensure that $q>0$ we have to take into account that $\tau/\lambda \leftarrow \Theta \leftarrow 2\tau/\lambda$. Then, the link to be used will be

$$\theta_i = \frac{(\tau/\lambda)\left[1 + \exp(\eta_i^T \beta)\right]}{1 + (1/(2\tau^2))\exp(\eta_i^T \beta)},$$

where β is a vector of regression coefficients and $(\eta 1, \ldots, \eta n)^T$, is the vector of covariates.

4. NUMERICAL ILLUSTRATIONS

Let us consider the automobile insurance portfolio used in de Jong and Heller (2008). This automobile insurance portfolio corresponds to the years 2004-05 and contains data of 67,856 insureds, of whom 4,624 filed a claim, i.e. a positive claims amount. Taking the claim amount as the dependent variable we have considered the following covariates: VALUEV: the vehicle value, in $10,000; VAG: the vehicle age (the reference category is the fourth); GENDER: the gender of the driver (0, male and 1 female); CLAIMS: the number of claims; VAGE: driver's age category (the reference category is the sixth).

Below in Table 1, parameter estimates, and *p*-values are shown when covariates are included in the model. It is noted that when the value of the threshold of *q=250* monetary units only the rate parameter λ and the covariates VAGE3 and VAGE4 are statistically significant at the usual significant levels, however, que this exceedance is increased, all the covariates are VALUEV and VAGE2 are statistically significant. The value of the negative of the log-likelihood function (NLL) is also shown for each value of the threshold.

Table 1. Parameter estimates and p-values when covariates are included in the model.

	$q = 250$		$q = 1000$	
Variable	Estimate	p-value	Estimate	p-value
VALUEV	0.128	0.639	0.536	0.339
VAG1	0.152	0.880	12.519	0.000
VAG2	0.059	0.943	12.967	0.000
VAG3	-0.154	0.860	-7.163	0.000
GENDER	4.691	0.550	8.114	0.000
CLAIMS	6.084	0.439	-7.392	0.000
VAGE1	-4.587	0.427	-2.905	0.000
VAGE2	3.333	0.541	-2.488	0.143
VAGE3	5.882	0.021	-8.821	0.000
VAGE4	7.128	0.016	-6.303	0.000
VAGE5	-14.047	0.383	-3.254	0.000
CONSTANT	-10.325	0.053	-9.380	0.000
λ	4.8E-4	0.000	3.3E-4	0.000
Observations	3843		2002	
NLL	33275.278		18084.285	

In Table 2 parameter estimates and *p*-values are shown when covariates and different inflation rates are included in the model. When the retention level is *q=250* and increasing the inflation rate from 1% to 2%, the number of statistically significant covariates at 5% level increases. However, the converse occurs when the retention level rises to *q=1000.* In this case, none of the covariates is statistically meaningful at the usual levels. Further analysis is required to obtain clearer conclusions on this matter.

Table 2. Parameter estimates and p-values when covariates and inflation are incorporated into the model.

	$q = 250$				$q = 1000$			
	Inflation = 1%		Inflation = 2%		Inflation = 1%		Inflation = 2%	
Variable	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
VALUEV	-0.001	0.994	-0.058	0.825	0.205	0.637	0.157	0.726
VAG1	0.966	0.438	1.865	0.363	4.056	0.617	2.997	0.439
VAG2	0.799	0.434	1.5[illegible]	0.434	4.231	0.588	3.146	0.423
VAG3	0.296	0.822	0.9[illegible]	0.671	-18.320	0.148	-74.385	0.260
GENDER	4.313	0.138	2.677	0.048	10.303	0.259	52.128	0.277
CLAIMS	5.002	0.128	2.973	0.025	-14.119	0.180	-46.866	0.265
VAGE1	-6.516	0.096	-3.193	0.008	-5.699	0.191	-18.716	0.267
VAGE2	1.899	0.046	12.911	0.061	23.279	0.171	74.954	0.253
VAGE3	6.340	0.134	-0.089	0.977	-15.842	0.149	-45.374	0.252
VAGE4	-4.432	0.063	-7.433	0.003	-14.587	0.153	-58.561	0.259
VAGE5	0.036	0.962	-1.212	0.004	-5.305	0.080	-19.663	0.242
CONSTANT	-5.777	0.076	-17.131	0.001	-18.332	0.185	-64.934	0.265
λ	4.8E-4	0.000	4.8E-4	0.000	3.3E-4	0.000	3.3E-4	0.000
Observations	3843		2002					
NLL	33281.021		33287.370		18084.774		18085.416	

REFERENCES

Albrecher, H., Beirlant, J., and Teugels, J. (2017). Reinsurance: Actuarial and Statistical Aspects.

de Jong, P., and Heller, G. (2008). Generalized Linear Models for Insurance Data. Cambridge University Press.

Dickson, D. (2005). Insurance Risk and Ruin. Cambridge University Press, Cambridge.

Hesselager, O. (1993). "A class of conjugate priors with applications to excess-of-loss reinsurance". ASTIN Bulletin, 23(1), 77–93.

Mata, A.J. (2000). "Pricing excess of loss reinsurance with reinstatements". ASTIN Bulletin, 30(2), 349–368.

APPENDIX

Estimation procedure by maximum likelihood and Fisher information matrices are provided in this for the model without covariates. Let us consider a random sample of n observations $z^{\sim} = (z_1, \ldots, z_n)$ from the distribution (2) and let $\Theta = (\lambda, q)$ be the vector of parameters to be estimated, assuming that τ is known. The log-likelihood function for Θ is proportional to

$$\ell(\tilde{z};\Theta) \propto n\left[2\log\lambda - \frac{\lambda\bar{z}}{\tau} - \log(1+\lambda q/\tau)\right] + \sum_{i=1}^{n}\log\big((z_i+q)/\tau\big)$$

where $\bar{z}$ is the sample mean.

The normal equations which provide the estimation of the parameters are given by

$$\frac{2}{\lambda} - \frac{\bar{z}}{\tau} - \frac{q}{\tau+\lambda q} = 0 \tag{4}$$

$$-\frac{n\lambda}{\tau+\lambda q} + \sum_{i=1}^{n}\frac{1}{z_i+q} = 0 \tag{5}$$

From (4) we get $q = \tau(2\tau - \bar{z}) / (\lambda(\lambda\bar{z} - \tau))$ which can be plugged into (5) to obtain an equation that depends only on the parameter λ and can be solved numerically. The second partial derivatives are as follows,

Considering that

$$\begin{aligned}\frac{\partial^2 \ell(\tilde{z};\Theta)}{\partial \lambda^2} &= -\frac{2n}{\lambda^2} + n\left(\frac{q}{\tau+\lambda q}\right)^2, \quad \frac{\partial^2 \ell(\tilde{z};\Theta)}{\partial q^2} = n\left(\frac{\lambda}{\tau+\lambda q}\right)^2 - \sum_{i=1}^{n} \frac{1}{(z_i+q)^2}, \\ \frac{\partial^2 \ell(\tilde{z};\Theta)}{\partial \lambda \partial q} &= -\frac{n\tau}{(\tau+\lambda q)^2}.\end{aligned}$$

where

$$E_Z[(z+q)^{-1}] = \varphi(\tau,\lambda,q) = \frac{\lambda^2\,\Gamma\left(0,\frac{q\lambda}{\tau}\right)}{\tau(\lambda q+\tau)} \exp\left(\frac{\lambda q}{\tau}\right),$$

$$\Gamma(a,m) = \int_m^\infty t^{a-1}\exp(-t)\,dt$$

is the incomplete gamma function, we have that the Fisher's information matrix can be approximated by

$$\mathcal{J}(\widehat{\Theta}) = \begin{bmatrix} \frac{2n}{\widehat{\lambda}^2} - n\left(\frac{\widehat{q}}{\tau+\widehat{\lambda}\widehat{q}}\right)^2 & \frac{n\tau}{(\tau+\widehat{\lambda}\widehat{q})^2} \\ \frac{n\tau}{(\tau+\widehat{\lambda}\widehat{q})^2} & -n\left(\frac{\widehat{\lambda}}{\tau+\widehat{\lambda}\widehat{q}}\right)^2 + \varphi(\tau,\widehat{\lambda},\widehat{q}) \end{bmatrix},$$

being $\widehat{\Theta} = [\widehat{\lambda},\widehat{q}]$the maximum likelihood estimators of the parameters. The asymptotic variancecovariance matrix of is obtained by inverting this information matrix.

MODELING EUROPEAN MORTALITY AT RETIREMENT AGE: A SPATIO-TEMPORAL ANALYSIS

Patricia Carracedo
Universitat Politècnica de València, Department of Applied Statistics, Operations Research and Quality, Alcoi, Spain
pcarracedo@eio.upv.es

Ana Debón
Universitat Politècnica de València, Centro de Gestión de la Calidad y del Cambio, Valencia ,Spain
andeau@eio.upv.es

ABSTRACT

During the 20th century, insurers and Social Security administrations focused on decreasing mortality in Europe. Despite these general trends, there are still significant differences in mortality levels between countries. Specifically, mortality modeling in both the demographic and actuarial fields is performed using country data. Thanks to this modeling, the variation in mortality between spatial units is captured. This study aims to complement this information using panel data models fitted by MATLAB and R software. The added value of these models is that they take into account the temporal and spatial dependence of mortality and identify the variables which influence it. The case study concerns the European retirement age's male and female mortality. The results confirm the spatial dependence between European countries and their neighbors during 1995–2012. Finally, we detail the similarities and differences found when using both software and the advantages and disadvantages of using each.

1. INTRODUCTION

The demographic dynamics of the aging population in developed European countries and other less developed countries are proposing reforms in welfare states and pensions in the European environment (Vaupel et al., 2012). Based on this, for the governments of European countries, it is essential to know if mortality is concentrated in identifiable geographic areas and, therefore, if an inequality gap is opening up. Spatial concentrations are produced by spatial dependence, which implies that the mortality of geographically close areas is more related than that of distant geographical areas (O'Hare & Li., 2014). That dependence is because neighboring countries may have similar social, economic, and cultural impacts.

Although we are aware of the significant role that research plays in health, focusing above all on people and time, in many cases, it does not consider the spatial and temporal dimensions of the data (Rezaeian et al., 2007). This fact is mainly due to the lack of adequate software to study data with a spatial and temporal dimension, that is, panel data (Millo and Piras, 2012). So far, according to our literature review, only the *splm* R-package (Millo and Piras, 2012), the code that complements Kapoor et al. (2007) in Stata, and finally, the code of Elhorst (2011) in MATLAB.

Therefore, the objective of this work lies in the ability to incorporate spatial dependence over time in mortality at advanced ages, and therefore, non-productive, of European countries over time through the model that best represents the location between countries using the MATLAB and R software.

Thus, this work aims to help public policies efficiently distribute resources and actuaries, who prepare life insurance and design pension plans. We intend to find synergies between different public administrations while transferring technology and knowledge. Thus, converting research into competitiveness helps decision-making between public administrations and insurers.

2. MATERIALS AND METHODS

2.1. Data

The database used comprises deaths and populations of 26 European countries from the Human Mortality Database (2016) for the period 1995–2012 with an age range of 65–110+ and male (m) and female (f) sexes. The available European countries are Austria, Belgium, Belarus, The Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, The Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland, The United Kingdom, and Ukraine. To explain the behavior of mortality as a function of socioeconomic variables, information about four variables for these 26 countries and 18 years was obtained from The World Bank Database (2021). These variables are the Gross Domestic Product (GDP), public health expenditure, CO2 emissions, and education expenditure based on a literature review of this area (Cutler et. al., 2006, Balan., 2016) and data availability for European countries from The World Bank Database (2021).

2.2. comparative mortality figure (CMF)

The Comparative Mortality Figure (CMF) was used in this study to compare the mortality experience over time by subpopulation and sex. The CMFs are the ratio between the number of expected deaths in the standard population if they had the age mortality rates of the subpopulations studied and the actual number of deaths in the standard population over a period (Julious et al., 2001):

$$\boldsymbol{CMF_{i,t,s}} = \frac{\boldsymbol{E_{i,t,s}}}{\boldsymbol{O_{2012,m}}} \quad i \in \{1,....,26\},\ t \in \{1995,....,2012\},\ s \in \{m, f\}$$

where $E_{i,t,s}$ represents the expected deaths in subpopulation *i*, in year *t*, and sex *s* and $O_{2012,m}$ is the number of observed deaths in the standard population in 2012 for males.

The CMF can be interpreted like this:

- CMF>1 indicates that there were more expected deaths than observed and a deaths excess exists.

- CMF ← 1 Indicates that more deaths were observed than expected, and a deaths deficit exists.

- CMF = 1 Indicates that there was the same number of expected deaths as observed.

2.3. Spatio-Temporal Panel Data Models

In a spatiotemporal panel, there may be dependency or correlation between the close observations (spatial units) over time (temporal units). Panel data consist of a cross-section of observations (individuals, countries, regions) followed through time. Specifically, the data from this study are panel data that combine a spatial dimension *N* (26 countries) and temporal dimension *T* (18 years).

Spatiotemporal panel data models are considered regression models that use the temporal and spatial heterogeneity of the panel structure to efficiently estimate parameters of interest (Elhorst., 2014). Unlike cross-sectional regression or time series, panel data models control for unobserved heterogeneity produced by spatial and temporal units. Additionally, they reduce multicollinearity problems between the variables by obtaining more efficient estimates of the parameters (Kennedy., 2003).

We propose a logarithmic spatiotemporal panel data model where both the CMF and the explanatory variables are logarithmically transformed to achieve approximate normality and symmetry in the CMF distribution and provide a straightforward interpretation of the results (Mendenhall et al., 1996).

The applied Spatio-temporal models are shown below:

- Spatial Lag Model (SLM);

$$\log (CMF_{it}) = \alpha + \lambda (I_T \otimes W_N) \log (CMF_{it}) + \log (X_{it})\, \beta + \in_{it}$$

- Spatial Lag Model with spatial fixed effects (SLMSFE);

$$\log (CMF_{it}) = \alpha + \lambda (I_T \otimes W_N) \log (CMF_{it}) + \log (X_{it}) \beta + \mu_i + \in_{it}$$

- Spatial Lag Model with time fixed effects (SLMTFE);

$$\log (CMF_{it}) = \alpha + \lambda (I_T \otimes W_N) \log (CMF_{it}) + \log (X_{it}) \beta + \nu_i + \in_{it}$$

- Spatial Lag Model with spatial and time fixed effects (SLMSTFE);

$$\log (CMF_{it}) = \alpha + \lambda (I_T \otimes W_N) \log (CMF_{it}) + \log (X_{it}) \beta + \mu_i + \nu_i + \in_{it}$$

- Log-log SLMSTFE adapted to Generalized Linear Model (GLM)

$$\log \left(\frac{E_{i,t,s}}{O_{2012,m}}\right) = \alpha + \lambda (I_T \otimes W_N) \log (CMF_{it}) + \log (X_{it}) \beta + \mu_i + \nu_i + \in_{it}$$

All these models are fitted by maximum likelihood with *splm* and *glm* functions in R software and the *sar_panel_FE* function in MATLAB. Highlight that the *glm* function considers the value of the quasiPoisson-likelihood and not the log-likelihood value.

3. RESULTS

Tables 1, 2, and 3 show the results of the spatiotemporal panel data models together with the corresponding goodness-of-fit measures by sex for *splm* function in R, *sar_panel_FE* function in MATLAB, and using the *glm* function in R, respectively.

According to the goodness-of-fit measures, the SLMSTFE has the highest R_2 (coefficient of determination) and the lowest σ_2 (residual deviance) in both software. However, the spatial effect in the model has more weight than the temporal effect, and the inclusion of the temporal effect increases both measures of goodness of fit, but not significantly. For this reason, and following the principle of parsimony, SLMSFE is considered the best model for both sexes. In this case, only the education expenditure variable is significant for SLMSFE in males with the glm function, simplifying the model.

Table 1. Model fittings using the splm function in R by sex.

Sex	Model	α	λ	log(GDP)	log(CO_2)	log(EE)	log(PHE)	logLik
Male	SLM							
	SLMSFE	1.436*	0.814 *	–0.038 *	0.054 *	0.050 *	0.034	145.081
	SLMTFE	1.677 *	0.420 *	–0.136 *	0.086 *	0.017	0.010	–213.404
	SLMSTFE	0.361	0.430 *	0.076 *	–0.024	0.038 *	0.031	249.972
Female	SLM							
	SLMSFE	2.826 *	0.544 *	–0.100 *	0.118 *	0.061 *	0.110 *	–32.906
	SLMTFE	1.486 *	0.374 *	–0.148 *	0.100 *	0.011	0.018	–343.423
	SLMSTFE	1.135 *	0.063	0.009	–0.001	0.040 *	0.071 *	49.788

*p-values < 0.05; Items in bold show the differences in the sign of the parameter estimation concerning Table 3.

Table 2. Model fittings using the sar_panel_FE function in MATLAB by sex.

Sex	Model	α	λ	log(GDP)	log(CO_2)	log(EE)	log(PHE)	R^2	logLik	σ_2
Male	SLM	1.592 *	0.460 *	–0.131 *	0.089 *	0.014	0.012	0.906	550.241	0.0051
	SLMSFE	1.479 *	0.806 *	-0,040 *	0.057 *	0.051 *	0.036	0.986	904.110	0.0008
	SLMTFE	1.680 *	0.419 *	–0.136 *	0.086 *	0.017	0.010	0.909	561.217	0.0051
	SLMSTFE	0.370	0.468 *	0.072 *	–0.023	0.038 *	0.030	0.988	1024.73	0.0007
Female	SLM	1.433 *	0.406 *	–0.144 *	0.102 *	0.008	0.020	0.851	424.616	0.0089
	SLMSFE	2.901 *	0.524 *	–0.103 *	0.123 *	0.062 *	0.114 *	0.964	735.632	0.0023
	SLMTFE	1.478 *	0.379 *	–0.147 *	0.100 *	0.011	0.018	0.855	431.176	0.0091
	SLMSTFE	1.148 *	0.108 *	0.007	–0.001	0.040 *	0.008 *	0.971	824.475	0.0019

*p-values < 0.05; Items in bold show the differences in the sign of the parameter estimation concerning Table 3

Table 3. Model fittings using the glm function in R by sex.

Sex	Model	α	λ	log(GDP)	log(CO_2)	log(EE)	log(PHE)	Residual Deviance
Male	SLM	1.346 *	0.540 *	–0.112	0.070*	0.004	0.017	3.371
	SLMSFE	0.490 *	0.982 *	–0.004	**–7.081×10^{-5}**	0.042 *	**–7.453×10^{-5}**	0.453
	SLMTFE	1.422 *	0.517 *	–0.117 *	0.070 *	**–**0.002	0.020	3.258
	SLMSTFE	0.356	0.760 *	0.037 *	–0.020	0.041 *	0.013	0.430
Female	SLM	1.157 *	0.505 *	–0.122 *	0.084 *	**–**0.008	0.029	3.494
	SLMSFE	1.932 *	0.760 *	–0.052 *	0.054 *	0.053 *	0.078 *	0.814
	SLMTFE	1.227 *	0.480 *	–0.128 *	0.083 *	**–**0.009	0.031 *	3.419
	SLMSTFE	1.161 *	0.122	0.012	–0.010	0.042 *	0.070 *	0.692

*p-values < 0.05; Items in bold show the differences in the sign of the parameter estimation concerning Tables 1 and 2

It is important to note that the corresponding R_2 for the OLS models by sex is 84.02% for males and 79.46% for females. These results indicate the importance of modeling the spatial correlation since including the model's spatial effect improves its fit by approximately 15% and 17% in males and females.

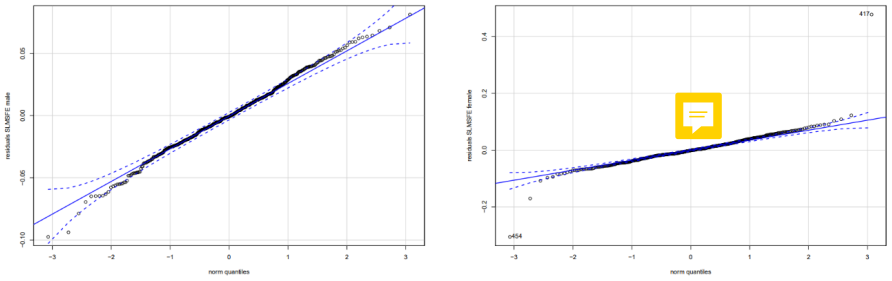
To test the homoscedasticity of the residuals of the SLMSFE model, the Breush-Pagan test (Breusch & Pagan., 1979) was applied. Table 4 shows the output of the Breusch-Pagan statistics and the corresponding p-values by gender. As the p-values of the test are greater than the 5% significance level for both sexes, there is not enough evidence to conclude that the residuals of the SLMSFE model are not homoscedastic.

Table 4. Values of the Breusch-Pagan statistic and p-values by sex

Model	Breusch-Pagan Test		*p*-Value	
	Male	Female	Male	Female
SLMSFE	42.407	34.300	0.103	0.358

Next, the SLMSFE model's residuals were analyzed for normality by sex using the quantile-quantile (Q-Q) plot. Figure 1 shows residuals are close to normally distribution for both sexes.

Figure 1. Normality of the SLMSFE model residuals for males and females.



(a) male (b) female

4. CONCLUSIONS

This study implements the Spatio-temporal panel models described in Section 2.3 using the *sar_panel_FE* function written by Elhorst (2011) in MATLAB and free software R using the *splm* and *glm* functions. Below is a summary of the differences and advantages found when using both software:

1. The main difference lies in the second or third decimal of the estimated parameters.

2. When the *glm* function is used in R, both the signs and the estimated parameters' values differ concerning the *splm* function in R and MATLAB. It could be because the *glm* function assumes a Poisson distribution, while *splm* and *sar_panel_FE* assume a normal distribution in maximum likelihood estimation.

3. Regarding the *splm* function, versions other than 1.3-7 and 1.5-2 do not correctly estimate the model parameters.

4. The goodness-of-fit measures in the output differ depending on the function used. The *splm* function only gives the log-likelihood value, the *sar_panel_FE* function gives the values for the coefficient of determination, the log-likelihood, and the residual variance, and the *glm* function gives the residual deviance value.

5. The log-likelihood values obtained in MATLAB are more reliable than in R because negative values appear in the latter. The three functions of the two software showed that the SLMSFE is the model that best fit produces.

6. An important advantage of the *glm* function compared to the rest is that the reference level for fixed effects can be changed. On the contrary, the *splm* and *sar_panel_FE* functions consider the reference level for fixed effects as the mean of fixed effects.

ACKNOWLEDGMENTS

The authors are grateful for financial support through a grant from Mapfre (ayudas a la investigación Ignacio H. de Larramendi 2017 Seguro y Previsión Social).

REFERENCES

Balan, F. (2016). "Environmental quality and its human health effects: A causal analysis for the EU-25". International Journal of Applied Economics, 13, 1, 57-71.

Breusch, T.S., and Pagan, A.R. (1979). "A simple test for heteroscedasticity and random coefficient variation". Econometrica: Journal of the Econometric Society, 1287-1294.

Cutler, D., Deaton, A., and Lleras-Muney, A. (2006). "The determinants of mortality". Journal of Economic Perspectives, 20(3), 97-120.

Elhorst, J.P. (2011). "Matlab software to estimate spatial panels". Version 2011-04-11, http://www.regroningen.nl/elhorst/software.shtml (routines downloaded on 6th May 2015).

Elhorst, J.P. (2014). "Spatial panel models". In *Handbook of Regional Science*; Springer: Berlin/Heidelberg, Germany; 1637-1652.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available online: www.mortality.org or www.humanmortality.de (accessed on 12 July 2016).

Julious, S.A., Nicholl, J., and George, S. (2001). "Why do we continue to use standardized mortality ratios for small area comparisons?". Journal of Public Health, *23*, 1, 40-46.

Kapoor, M., Kelejian, H.H., and Prucha I.R. (2007). "Panel data models with spatially correlated error components". Journal of Econometrics, 140, 1, 97–130.

Kennedy, P. (2003). A guide to econometrics Cambridge. MIT Press: Cambridge, MA, USA

Mendenhall, W., Sincich, T., and Boudreau, N.S. (1996). A Second Course in Statistics: Regression Analysis. Prentice Hall: Upper Saddle River, NJ, USA, 5.

Millo, G., and Piras G. (2012). "splm: Spatial panel data models in R". Journal of Statistical Software, 47.1. http://www.jstatsoft.org/v47/i01/, 1-38.

O'Hare, C., and Li, Y. (2014). "Is mortality spatial or social?". Economic Modelling, 42, 198-207.

Rezaeian, M., Dunn, G., and St Leger, S. (2007). "Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary". Journal of Epidemiology & Community Health, 61, 98-102.

The World Bank Database. World Development Indicators; The World Bank Database: Washington, DC, USA, 2018. Data Download on 6 March 2018. Available online: https://data.worldbank.org/indicator (accessed on 9 February 2021).

Vaupel, J.W., Zhang, Z., and van Raalte, A.A. (2011). "Life expectancy and disparity: an international comparison of life table data". BMJ Open, 1, 1.

COMPARACIÓN DE RIESGOS ELEVADOS EN LA TEORÍA DUAL DE YAARI

Antonia Castaño-Martínez
Universidad de Cádiz, Departamento de Estadística e I.O., España
antonia.castano@uca.es

Gema Pigueiras
Universidad de Cádiz, Departamento de Estadística e I.O., España
gema.pigueiras@uca.es

Miguel A. Sordo
Universidad de Cádiz, Departamento de Estadística e I.O., España
mangel.sordo@uca.es

Carmen D. Ramos
Universidad de Cádiz, Departamento de Estadística e I.O., España
carmen.ramos@uca.es

ABSTRACT

En el contexto de la Teoría de la Utilidad, Yaari (1987) propuso una teoría dual de elección bajo riesgo según la cual, la aversión al riesgo viene caracterizada por una función (llamada distorsión) que modifica la probabilidad del riesgo de cola. Bajo esta propuesta, Wang y Young (1998) obtienen una clase de órdenes parciales para la valoración de riesgos que caracterizan a través de una clase de distorsiones. Sin embargo, bajo este enfoque, dados dos riesgos X e Y tales que X es menos arriesgado que Y, entonces sus valores medios siempre están ordenados en la misma dirección. Esto puede ser inadecuado cuando la aseguradora está más preocupada por los riesgos más altos que

por los medios ya que, si bien los eventos de cola son raros, pueden tener importantes consecuencias negativas. En este trabajo se obtiene una clase de órdenes más débiles de valoración de riesgos, que caracterizan la aversión a los riesgos más elevados y bajo los cuales, las medias no tienen por qué estar necesariamente ordenadas en la misma dirección.

1. INTRODUCCIÓN

En el contexto de la teoría dual de elección bajo riesgo de Yaari (1987), dado un riesgo X con función de distribución $F(x)$ y función de supervivencia $\overline{F}(x) = 1 - F(x)$, un agente de seguros utiliza una distorsión g (i.e., una función no decreciente de [0,1] en [0,1] tal que $g(0) = 0$ y $g(1) = 1$ que modifica la función de supervivencia y evalúa el riesgo X del siguiente modo

$$H_g(X) = \int_0^\infty g\big(\overline{F}(x)\big)dx. \qquad (1)$$

En el contexto de los principios de prima de Wang (1996), $H_g(X)$ representa el precio (prima) por aceptar el riesgo X. Un requerimiento fundamental de las aseguradoras es cobrar al menos la pérdida esperada $E(X)$ (i.e., la prima neta que se obtiene con la distorsión $g(t) = t$, para $0 \leq 0 \leq 1$), lo cual queda garantizado cuando la función g es cóncava. En este caso además se satisfacen otras propiedades deseables y, de hecho, el principio de prima resulta ser coherente (ver Younq, 2004). Un ejemplo de principio de prima distorsionado, con distorsión $g(t) = \min\{\frac{t}{1-p}, 1\}$, es el riesgo de valor en cola ($TVaR$) dado por

$$TVaR_p(X) = \frac{1}{1-p}\int_p^1 F^{-1}(t)dt, \;\; p \in (0,1),$$

donde $F^{-1}(t) = \inf\{x: F(x) \geq t\}$, $0 \leq t \leq 1$.

Si estamos interesados en ordenar un conjunto de agentes según su aversión al riesgo, podríamos clasificarlos según ciertas características de las distorsiones utilizadas. Puesto que si una distorsión g es cóncava, siempre proporciona primas mayores o iguales que la prima neta $E(X)$, podemos caracterizar la aversión al riesgo según la concavidad de g. Wang y Young (1998) consideran una secuencia de distorsiones progresivamente más aversas al riesgo.

Definición 1. Dado $n \geq 1$, D_n es la clase de distorsiones g tales que g es al menos $n-1$ veces diferenciable, $(-1)^{k+1} g^{(k)} \geq 0$ para $k = 1,\ldots, n-1$ y $(-1)^n g^{(n-1)}$ es no creciente.

Observemos que $D_{n+1} \subset D_n$ para $n = 1,2,\ldots$ y que en la secuencia $\{D_n, n \geq 1\}$, las distorsiones son más aversas al riesgo a medida que crece n. La Definición 1 induce el siguiente orden.

Definición 2. Dados dos riesgos X e Y, $X \leq_n Y$ si y solo si $H_g(X) \leq H_g(Y)$ para todo $g \in D_n$.

La relación $\leq_n$ puede ser caracterizada por un orden basado en las integrales sucesivas de la función cuantil. Denotemos ${}^1F_X^{-1}(p) = F_X^{-1}(p)$ y

$$ {}^{n+1}F_X^{-1}(p) = \int_p^1 [\, {}^nF_X^{-1}(t)]dt, \;\; n = 1,2,\ldots, \;\; 0 \leq p \leq 1. $$

Definición 3. Dados dos riesgos X e Y con funciones de distribución F_x y F_y, respectivamente, X es menor que Y en el orden estocástico n-ésimo dual (denotado por $X \leq_{n-dual} Y$) si ${}^nF_X^{-1}(p) \leq {}^nF_Y^{-1}(p)$ para todo $p \in (0,1]$ y $E[max(X_1,\ldots,X_k)] \leq [Emax(Y_1,\ldots,Y_k)]$ para $k = 1,\ldots, n-1$ donde $X_1,\ldots,X_k$ e $Y_1,\ldots,Y_k$ son copias independientes de X e Y, respectivamente.

Para $n = 1$, el orden estocástico n-ésimo dual coincide con el orden estocástico usual ($\leq_{st}$) y para $n = 2$ con el orden stop-loss ($\leq_{sl}$). Wang y Young (1998) probaron el siguiente resultado.

Teorema 4. Dado $n = 1, 2,\ldots, H_g(X) \leq H_g(Y)$ para todo $g \in D_n$ si y solo si $X \leq_{n-dual} Y$.

El Teorema 4 proporciona una motivación, tanto desde el punto de vista económico como desde el estadístico, para la secuencia de órdenes dada en la Definición 3. Sin embargo, bajo este enfoque, dados dos riesgos X e Y tales que X es menos arriesgado que Y, entonces sus valores medios siempre están ordenados en la misma dirección. Esto puede ser inadecuado cuando la aseguradora está más preocupada por los riesgos más altos que por los medios ya que, si bien los eventos de cola son raros, pueden tener importantes consecuencias negativas.

Ejemplo 5. Consideremos *X* ~ *Pareto* (2.4,3) e *Y* ~ *Pareto* (2.2). En este caso claramente $X \not\leq_{n\text{-}dual} Y$ ya que $E(X) = 2.143 > 2 = E(Y)$. Por otro lado, si consideramos la distorsión $g_1(t) \in D_n$, para $n \geq 1$, dada por

$$g_1(t) = \begin{cases} t(1 - \log t), & 0 < t \leq 1 \\ 0, & t = 0, \end{cases}$$

se obtiene $H_{g1}(X) = 5.8163 \leq 6 = H_{g1}(Y)$, por lo que alguien interesando en evitar grandes pérdidas, percibirá el riesgo *Y* más arriesgado que el *X*. $H_{g1}(X)$ (que de ahora en adelante denotaremos por $I_1(X)$) posee interpretaciones interesantes para los actuarios como, por ejemplo

$$I_1(X) = \int_0^1 TVaR_p(X)dp$$

(ver Sordo et al., 2016).

En la Sección 2 introduciremos una secuencia de subclases $\widehat{D}_n \subset D_n$ con el objetivo de estudiar la secuencia de ordenamientos de riesgos basados en esperanzas distorsionadas de la forma (1). Puesto que cualquier medida de riesgo de la forma (1) con distorsión asociada cóncava puede ser escrita como mixturas de *TVaRs* de la siguiente forma

$$I_h(X) = \int_0^1 TVaR_p(X)dh(p) \qquad (2)$$

siendo *h* una función de ponderación[5] (es decir, una función no decreciente de [0,1] en [0,1] tal que h(0) = 0 y h(1) = 1) (ver Pflug y Römisch, 2007). Se obtendrá una secuencia basada en medidas de la forma (2) con correspondencia uno a uno con la secuencia de órdenes anteriormente obtenida. En la Sección 3, introducimos un criterio de dominancia estocástica basado en las integrales sucesivas del *TVaR* y que resulta ser equivalente a la secuencia de ordenaciones introducida en la Sección 2. En la Sección 4 finalmente tenemos las conclusiones.

[5] Aunque *h* es definida formalmente como una distorsión, en este trabajo preferimos no llamarla así ya que *h* no está distorsionando directamente a la función de supervivencia *F*.

A lo largo de este artículo, dada una función *f*, $f^{(k)}$ denotará su derivada *k*-ésima, *k* = 1,2, ... , $f^{(0)} = f$.

2. DOS SECUENCIAS DE ORDENACIÓN DE RIESGOS

En esta sección consideraremos dos secuencias de ordenamiento de riesgos basadas en dos tipos diferentes de funciones. La primera se basa en medidas de riesgo distorsionadas de la forma (1) y la segunda en medidas de riesgo de la forma (2).

Definición 6. Para $n \geq 3$, $\widehat{D}_n$ es la clase de distorsiones *g* tales que es al menos *n* veces diferenciable, $g'(1) = 0$,

$$r_{g,k}(t) = \frac{-tg^{(k+1)}(t)}{g^{(k)}(t)} \geq k-1, \quad k = 1, \ldots, n-2,$$

Y $(-1)^n \left(tg^{(n-1)}(t) + (n-3)\, g^{(n-2)}(t)\right)$ is non-increasing in *t*.

Observación 7. Observemos que $\widehat{D}_n \subset D_n$, para $n \geq 3$.

Definición 8. Para $n \geq 1$, C_n es la clase de funciones de ponderación *h* tales que *h* es al menos *n-1* veces diferenciable, $h^{(k)} \geq 0$ para $k = 1,\ldots,n-1$ y $h^{(n-1)}$ es no decreciente.

El siguiente resultado muestra que un para un decisor, el riesgo evaluado por (2) usando $h \in C_n$ es equivalente al evaluado por (1) con $g \in \widehat{D}_{n+1}$.

Teorema 9. Dadas *X* e *Y* dos variables aleatorias no negativas y dado $n \geq 2$, entonces $I_h(X) \leq I_h(Y)$ para todo $h \in C_n$ si y solo si $H_g(X) \leq H_g(Y)$ para todo $g \in \widehat{D}_{n+1}$.

El siguiente Corolario muestra que un agente que evalúe el riesgo con una distorsión en $\widehat{D}_n$ es más averso al riesgo que un agente que evalúe el riesgo con una distorsión en $\widehat{D}_n$ (pero no en $\widehat{D}_n$).

Corolario 10. Sea $g \in \widehat{D}_n$, $n \geq 3$. Entonces $H_g(X) \geq I_1(X)$.

Ejemplo 11. Para $m \geq 1$ y $n \geq 2$, $h(t) = t^m \in Cn$. En este caso la distorsión g correspondiente viene dada por

$$g_m(t) = mt \int_t^1 \frac{(1-u)^{m-1}}{u} du + 1 - (1-t)^m, \qquad 0 < t \leq 1,$$

pertenece a $\widehat{D}_n$ para $n \geq 3$. La correspondiente medida de riesgo distorsionada es

$$I_m(X) = \int_0^1 TVaR_t(X) dt^m = E[X|X > max\{X_1, \dots, X_n\}], \tag{3}$$

con $X_1, \dots, X_n$ copias independientes de X.

3. DOMINANCIA ESTOCÁSTICA

Denotemos $T_X^{[1]}(p) = TVaR_p(X)$ y definamos

$$T_X^{[n]} = \int_p^1 T_X^{[n-1]}(t) dt, \;\; n = 2,3, \dots, 0 \leq p \leq 1.$$

Sabemos (ver Lema 2.1 en Sordo y Ramos, 2007) que $X \leq_{sl} Y$ si y solo si $T_X^{[1]}(p) \leq T_Y^{[1]}(p)$ para todo $p \in [0,1]$. En este caso podemos utilizar $T_X^{[n]}, n = 2,3, \dots$ y la clase de medidas $I_m(X)$ dadas por (3) para definir la siguiente secuencia de ordenaciones de riesgos.

Definición 12. Dados dos riesgos X e Y no negativos y dado $n \geq 2$, decimos que X es menor que Y en el n-esimo orden $TVaR$ (y lo denotamos $X \leq_{tavr[n]} Y$) si y sólo si $T_X^{[n]}(p) \leq T_Y^{[n]}(p)$ para todo $p \in [0,1]$ e $I_k(X) \leq I_k(Y)$ para $k = 1, \dots, n-1$.

Según el Teorema 4.4 de Wang y Young (1998), el orden stop-loss es equivalente a la ordenación según la clase de índices $\{H_g(X), g \in D_2\}$. En esta sección probamos que para $n \geq 2$ el orden $\leq_{tavr[n]}$ es equivalente a la ordenación según la clase de índices $\{H_g(X), g \in \widehat{D}_{n+1}\}$.

Teorema 13. Dados dos riesgos X e Y no negativos y dado $n \geq 2$, entonces $X \leq_{tavr[n]} Y$ si y sólo si $H_g(X) \leq H_g(Y)$ para todo $g \in \widehat{D}_{n+1}$.

Ejemplo 14. Retomando el Ejemplo 5 con $X \sim$ *Pareto* (2.4,3) e $Y \sim$ *Pareto* (2.2), donde vimos que $X \not\leq_{n-dual} Y$, se obtiene que $X \leq_{tavr[2]} Y$ y por tanto, X es menos arriesgada que Y para aquellos aseguradores con una distorsión $g \in \widehat{D}_3$..

4. CONCLUSIONES

Se ha definido una secuencia de subclases $\widehat{D}_n \subset D_n$ con el objetivo de ordenar una secuencia de riesgos basados en esperanzas distorsionadas. Debido a la relación existente entre las medidas de riesgo distorsionadas con distorsiones cóncavas y las medidas de riesgo que vienen dadas como áreas ponderadas bajo la curva *TVaR*, se ha definido otra secuencia de subclases basada en funciones de ponderación (C_n). Se ha establecido una correspondencia uno a uno entre las distorsiones en $\widehat{D}_{n+1}$ y las funciones de ponderación en C_n. Finalmente, para aquellos agentes más preocupados por los riesgos más altos que por los medios, se ha introducido una secuencia de criterios de dominancia estocástica basado en las integrales iteradas de *TVaRs*, la cual resulta ser equivalente a la secuencia de ordenación de riesgos dada por las clases $\widehat{D}_n$.

AGRADECIMIENTOS

Esta investigación ha sido parcialmente financiada por el Ministerio de Economía y Competitividad (España) bajo la subvención MTM2013-46962-C2-2-P y por el 2014–2020 ERDF Programa Operativo y el Departamento de Economía, Conocimiento, Empresa y Universidad del Gobierno Regional de Andalucía bajo la subvención FEDER-UCA18-107519.

BIBLIOGRAFÍA

Pflug, G.C., and Römisch, W. (2007). Modeling, Measuring and Managing Risk, World Scientific Books.

Sordo, M.A., Castaño-Martínez, A., and Pigueiras, G. (2016). "A family of premium principles based on mixtures of TVaRs". Insurance, Mathematics and Economics, 70, 397-405.

Sordo, M.A., and Ramos, H. (2007). "Characterization of stochastic orders by L-functionals". Statistical Papers, 48, 249-263.

Wang, S. (1996). "Premium calculation by transforming the layer premium density". Astin Bulletin, 26, 71-92.

Wang, S., and Young, V.R. (1998). "Ordering risks: Expected utility theory versus Yaari's dual theory of risk". Insurance, Mathematics and Economics, 22, 145-161.

Yaari, M.E. (1986). "Univariate and multivariate comparisons of risk aversion: A new approach". Cambridge University Press.

Yaari, M.E. (1987). "The dual theory of choice under risk". Econometrica 55, 95-115.

Young, V.R. (2004). "Premium Principles". In: Encyclopedia of Actuarial Science, Wiley, New York.

ALTERNATIVE SCORING FUNCTION SPECIFICATIONS FOR ESTIMATING VALUE AT RISK AND CONDITIONAL TAIL EXPECTATION

Xenxo Vidal-Llana
Universitat de Barcelona, Departament d'Econometria, Estadística i Economia Aplicada, Barcelona, Spain
juanjose.vidal@ub.edu

Vincenzo Coia
University of British Columbia, Department of Statistics, Vancouver, BC Canada

Montserrat Guillen
Universitat de Barcelona, Departament d'Econometria, Estadística i Economia Aplicada, RISKcenter-IREA, Barcelona, Spain
mguillen@ub.edu

ABSTRACT

Choosing a suitable scoring function to fit Conditional Tail Expectation regression that depends on covariates has been in the discussion since the proposal of the general specification. We use a Non-Crossing Dual Neural Network, a model that can predict Value at Risk and Conditional Tail Expectation for several levels with non-crossing conditions, and we apply it to quantile levels 0.9 and 0.99. We show alternative score function specifications and discuss what appears to be the best choice to the analyst. We examine examples of implementation in the domain of risk analysis, in particular in risk measurement for telematics driving data.

1. INTRODUCTION

Algorithms such as Quantile Regression (Koenker and Basset Jr. (1978)) provided a breath of fresh air in the analysis of heavy tailed and asymmetric distributions of dependent variables. For a risk analyst, the ability to identify characteristics that have a significant effect on the occurrence of strange and costly events is essential.

Risk analysts use Value at Risk (VaR) as one of the main drivers for their decision making and asset management, but, in the recent years, they are becoming increasingly interested on the Conditional Tail Expectation (CTE) because of its ability to summarize the behavior of a tail and because CTE is behind the motivation proposed by the Basel III agreement (Basel Committee of Banking Supervision (2016)).

While a change from VaR to CTE seems a step in the right direction, it possesses a big drawback, CTE is not elicitable. Elicitability is defined as the existence of a consistent scoring function (Gneiting (2011)) to derive a CTE estimator, in other words, the capacity that CTE is calculated directly by optimizing a loss function does not exist. Clearly, the CTE depends on the VaR to be calculated beforehand, so it is not an elicitable risk measure. But the pair (VaR, CTE) is elicitable (Fissler and Ziegel (2016)), making possible their joint estimation. Those authors propose a scoring function for the pair (VaR, CTE) that, in its general form, depends on two subfunctions that must meet specific properties to ensure consistency of the predictions.

Another line of research related to this one is the non-crossing quantile models. The non-crossing property means that the CTE estimate is beyond is corresponding VaR value, i.e., for losses expressed as positive numbers and right tails, the CTE should be equal or larger that VaR, and VaR estimates should be increasing with increasing quantile levels, i.e., VaR at the 0.90 level would be smaller than VaR at the 0.99 level. First studied by He (1997) and Yu et al. (2003), those authors motivate the calculation of different quantile levels while solving a native problem in the scoring function, the crossing of the predictions. There are also improvements among the last years in which a grid of quantile levels can be used to estimate different VaR without crossing using Neural Networks (Cannon (2018); Moon (2021)). Related to the literature of the CTE estimation, Acerbi and Szekely (2014)

proposed a method to assess the non-crossing between the VaR and CTE but did not open the debate about several quantile levels.

There are several fields in which a different number of quantile levels and their VaR and CTE must be calculated jointly, like telematics in insurance pricing, constituting reserves in a bank or insurance company, extreme events in weather forecast, or flood prediction using riverbed widths. That is why we are going to use the Non-Crossing Dual Neural Network, proposed by Vidal-Llana et al. (2022), which can calculate VaR and CTE for different quantile levels while meeting certain non-crossing properties. Our study relies on the comparison between the choice of subfunctions of the scoring function, in order to select which subfunction in the score is more adequate for the risk analyst. We apply this methodology to a telematics dataset, to simulate the study that an insurance company could do to predict driving behaviors of policyholders based on their driving skills.

This manuscript is organized as follows. We present the methodology used in Section 2, an introduction to the dataset used in Section 3, the main results on Section 4 and we conclude in Section 5.

2. METHODOLOGY

Calculating the Value at Risk (VaR) corresponds to predicting the quantile of the conditional distribution of the response variable, as a function of some covariates. For a random variable Y with probability distribution function F_Y, the VaR for the quantile level q is defined as:

$$VaR_q(Y) = \inf\{y \in \mathbb{R} \mid F_Y(y) > q\} = F_Y^{-1}(1-q)\,.$$

Proposed by Koenker and Basset Jr. (1978), the scoring function for calculating the quantile level q of an observed variable y is the following:

$$\rho_q(r_1, y) = (q - \mathbb{I}_{\{y-r_1<0\}})(y - r_1)$$

where $\mathbb{I}$ is the identification function, with value 1 when the subscript is met, and 0 otherwise.

Moving from the VaR to the Conditional Tail Expectation (CTE) has a main motivation: the quantile does not provide information about the behavior of the tail, so if we are interested on extreme events, evaluating the CTE becomes a crucial part of risk assessment. The CTE for a quantile level q for the right part of the tail is defined as:

$$CTE_q(Y) = \mathbb{E}[Y \mid Y \geq VaR_q(Y)]$$

Proposed by Fissler and Ziegel (2016), a consistent scoring function for estimating both VaR and CTE together has the following form:

$$S_q(r_1, r_2, y) = \mathbb{I}_{\{y>r_1\}}\big(-G_1(r_1) + G_1(y) - G_2(r_2)(r_1 - y)\big) + (1-q)\big(G_1(r_1) - G_2(r_2)(r_2 - r_1) + \mathcal{G}_2(r_2)\big)$$

S_q is a consistent scoring function if G_1 is an increasing function, G_2 is increasing and concave, and $\mathcal{G}_2' = G_2$.

In this study, we focus on the different choices of functions G_1 and G_2 shown in Table 1. Some of the choices have already been presented in the literature, for example LF_0 was proposed by Taylor (2019) and LF_3 by Fissler and Ziegel (2016). The others are modifications of the previous ones or our own proposals that meet the criteria for creating a consistent scoring function.

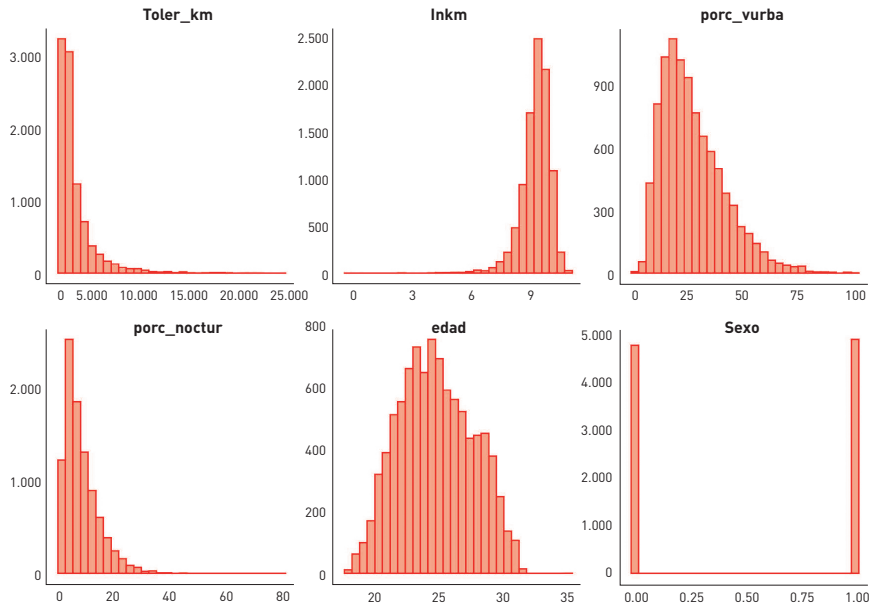
Table 1. Choices of G_1 and G_2 used in this study.

	$G_1(z)$	$G_2(z)$
LF_0	0	$\frac{1}{z}$
LF_1	z	$\frac{1}{z}$
LF_2	0	$\frac{e^z}{(e^z+1)^2}$
LF_3	z	$\frac{e^z}{(e^z+1)^2}$
LF_4	z	1
LF_5	z	e^{-z}
LF_6	z	$\frac{1}{e^z+1}$

3. DATASET

The data set used in the empirical part of this manuscript corresponds to information collected from drivers during one year, which contains variables such as kilometers traveled in logarithms (*lnkm*), the percentage distance driven in urban areas (*porc_vurba*), the percentage of night driving (*porc_noctur*), age (*edad*), and gender (*sexo*). The number of observations treated in our study is 9,614. Our main objective is predicting the amount of km that have been driven above the speed limit (*toler_km*). We plot the distribution of the response variable and covariates in Figure 1.

Figure 1. Distribution of the response variable and the covariates.



In Figure 1 we observe the distribution of the response variable (top-left panel) and the covariates. The distributions show the presence of heavy tails, for example in the response (*toler_km*) and other covariates, like *lnkm* (with left tail instead of right), *porc_vurba* and *porc_noctur*, meaning a possibility of high-risk events to happen, which the insurance company must keep in mind when evaluating future

risks, creating a pay-as-you-go scheme for pricing their policyholders, or constituting reserves. These data were also used by Pitarque and Guillen (2022) and Guillen et al. (2019, 2021b). Guillen et al. (2021a) show the use of quantile-charts to identify risky drivers.

4. RESULTS

We predict Value at Risk (VaR) and Conditional Tail Expectation (CTE) for quantile levels 0.7, 0.8, 0.9, 0.925, 0.95, 0.975 and 0.99, but we will only report 0.9 and 0.99 for simplicity. We present results obtained by running 20 different seeds with each of the 7 loss function specifications presented in Table 1. We choose the best seed for each model by normalizing and adding losses of all quantiles predicted across the whole dataset.

Figure 2. Computational time to run each model.

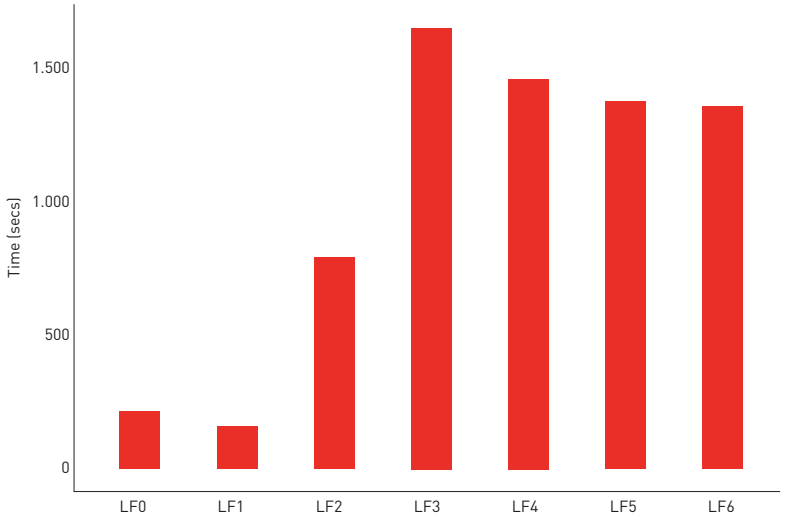


Table 2. Normalized scoring for each model (rows) using each scoring function (columns). Lower value means better prediction. In bold the lower value of each column, which represents the best model scored with the column's loss function.

		Evaluated with						
	q = 0.9	LF_0	LF_1	LF_2	LF_3	LF_4	LF_5	LF_6
Estimated with	LF_0	0,501993	0,990087	0,511483	0,990453	0,999898	0,990008	0,990008
	LF_1	0,501976	0,988420	0,514710	0,988989	0,999861	0,988329	0,988329
	LF_2	0,501908	0,978600	0,492735	0,977818	0,999515	0,978440	0,978440
	LF_3	0,501910	0,978595	0,508696	0,979157	0,999514	0,978435	0,978435
	LF_4	0,502239	0,978605	0,513918	0,988393	0,999514	0,978416	0,978416
	LF_5	0,501909	0,978266	0,497490	0,977887	0,999499	0,978103	0,978103
	LF_6	0,501906	0,977538	0,506271	0,977921	0,999464	1,000000	0,977377

		Evaluated with						
	q = 0.99	LF_0	LF_1	LF_2	LF_3	LF_4	LF_5	LF_6
Estimated with	LF_0	0,500264	0,886228	0,599210	0,920846	0,983752	0,886122	0,886122
	LF_1	0,500274	0,890705	0,603169	0,925225	0,985134	0,890598	0,890598
	LF_2	0,500234	0,751875	0,544446	0,783462	0,901624	0,751700	0,751700
	LF_3	0,500250	0,777430	0,554831	0,813052	0,924107	0,777257	0,777257
	LF_4	0,500231	0,770178	0,553829	0,818493	0,918098	0,770014	0,770014
	LF_5	0,500238	0,772509	0,552066	0,806996	0,920060	0,772342	0,772342
	LF_6	0,500249	0,769042	0,551959	0,803873	0,917145	1,000000	0,768927

In Figure 2 we show the computational time (in seconds) that each 20 seeds have lasted to run for each model specification. We see how LF_0 and LF_1, the simplest specifications, take a low amount of computer running time in comparison of the other scoring functions, around 200 seconds. LF_2 presents a medium time with approximately 800 seconds to run, and the rest take a similar amount of time, around 1500 seconds.

We present the results of the normalized scores for each model in Table 2. In this table, we score each model's predictions with each loss function from Table 1. We observe two main outputs from this table. The scoring functions with specifications LF_0 and LF_1 do not give a good result in comparison to the other models, as

their corresponding rows present significantly higher values than the others. Secondly, LF_2 and LF_6 are the two best performing specifications across all scoring functions used to evaluate, as they present the lowest scores of each column. We decide that the best performing one is LF_2 because it halves the computational time to run the experiments in comparison to LF_6, as shown in Figure 2. Intermediate quantile levels also provide a better score for LF_2.

5. CONCLUSIONS

When evaluating heavy tailed distributions and having to predict both Value at Risk (VaR) and Conditional Tail Expectation (CTE) for several quantile levels, analysts need to make choices before defining the scoring function to estimate the model. This is an additional problem to the analyst. In this manuscript, we enlighten this area showing the results of several specifications and implement a method to compare scoring choices.

We have estimated VaR and CTE for a telematics dataset for quantile levels of the right part of the tail using a Neural Network based model that supports VaR and CTE non-crossing conditions. We show results for quantile levels 0.9 and 0.99 to mimic the study that an insurance company could do to evaluate high risk profiles among its policyholders.

Results suggest that the LF_2 specification for the general scoring function of the pair (VaR, CTE), with $G_1(z) = 0$ and $G_2(z) = \frac{e^z}{(e^z+1)^2}$, gives the best approximation to the quantiles, as it shows better estimation performance and a medium computer time consumption.

Two natural continuations of this article would be: 1) to study, within the telematics landscape, the left part of the tail to identify low risk drivers, i.e., customers with low probability of filing a claim, and 2) to show if the same conclusions hold in different datasets, and to see whether or not the choice of LF_2 is the best performing one, regardless of the distribution of the response variable.

ACKNOWLEDGMENTS

We want to have a special acknowledgement to Fundació Banc Sabadell: "Ajudes a la investigació 2022", Fundación BBVA: "Ayudas a proyectos de investigación en Big Data", AGAUR: "PANDÈMIES" grant and the Spanish Ministry of Science grant PID2019-105986GB-C21 for their support to our research.

REFERENCES

Basel Committee on Banking Supervision (2016). Minimum capital requirements for market risk.

Cannon, A.J. (2018). "Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes". Stochastic Environmental Research and Risk Assessment, 32, 11, 3207-3225.

Fissler, T., and Ziegel, J.F. (2016). "Higher order elicitability and Osband's principle". The Annals of Statistics, 44, 4, 1680-1707.

Gneiting, T. (2011). "Making and evaluating point forecasts". Journal of the American Statistical Association, 106, 494, 746-762.

Guillen, M., Bermúdez, L., and Pitarque, A. (2021). "Joint generalized quantile and conditional tail expectation regression for insurance risk analysis". Insurance: Mathematics and Economics, 99, 1-8.

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

Guillen, M., Pérez-Marín, A.M., and Alcañiz, M. (2021). "Percentile charts for speeding based on telematics information". Accident Analysis & Prevention, 150, 105865.

He, X. (1997). "Quantile curves without crossing". The American Statistician, 51, 2, 186-192.

Koenker, R., and Bassett Jr, G. (1978). "Regression quantiles". Econometrica: Journal of the Econometric Society, 46, 1, 33-50.

Moon, S.J., Jeon, J.J., Lee, J.S.H., and Kim, Y. (2021). "Learning multiple quantiles with neural networks". Journal of Computational and Graphical Statistics, 30, 4, 1238-1248.

Pitarque, A., and Guillen, M. (2022). "Interpolation of quantile regression to estimate driver's risk of traffic accident based on excess speed". Risks, 10, 1, 19.

Taylor, J.W. (2019). "Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution". Journal of Business & Economic Statistics, 37, 1, 121-133.

Vidal-Llana, X., Salort Sanchez, C., Coia, V., and Guillen, M. (2022). "Non-Crossing Dual Neural Network: Joint Value at Risk and Conditional Tail Expectation estimations with non-crossing conditions". IREA-Working Papers, 2022, IR22/12.

Yu, K., Lu, Z., and Stander, J. (2003). "Quantile regression: applications and current research areas". Journal of the Royal Statistical Society: Series D (The Statistician), 52, 3, 331-350.

LA REFORMA DEL CÁLCULO DE LA BASE REGULADORA Y SU IMPACTO EN LA SOSTENIBILIDAD DEL SISTEMA DE PENSIONES ESPAÑOL

Enrique Devesa Carpio
Universidad de Valencia, Departamento Economía Financiera y Actuarial, España
Enrique.Devesa@uv.es

Mar Devesa Carpio
Universidad de Valencia, Departamento Economía Financiera y Actuarial, España
Mar.Devesa@uv.es,

Inmaculada Domínguez Fabián
Universidad de Extremadura, Departamento Economía Financiera y Contabilidad, España
idomingu@unex.es

Borja Encinas Goenechea
Universidad de Extremadura, Departamento Economía Financiera y Contabilidad, España
bencinas@unex.es

Robert Meneu Gaya
Universidad de Valencia, Departamento Matemàtiques per a l'Economia i l'Empresa, España
Robert.Meneu@uv.es

ABSTRACT

Los dos componentes principales a la hora de calcular la pensión inicial de jubilación son el porcentaje por años cotizados y la Base Reguladora. En este trabajo se ha cuantificado el impacto que puede tener la modificación del

cálculo de la Base Reguladora sobre la cuantía de la pensión inicial y sobre la sostenibilidad del sistema de pensiones español cuantificando el ahorro/gasto en pensiones que se generaría tanto en términos de caja como de valor actual actuarial. La multiplicidad de definiciones que se pueden aplicar a la Base Reguladora complica su evaluación. Nos vamos a centrar en la ampliación del periodo de cálculo desde los 25 años actuales a 35, con el añadido de la posible elección de los 25 mejores años, dentro de los últimos 35. Para ello se ha utilizado la Muestra Continua de Vidas Laborales (MCVL), aplicando la metodología actuarial. Los resultados demuestran que el aumento de 25 a 35 años provocaría una disminución de un 8,7% de la cuantía de la pensión inicial y una mejora de la sostenibilidad del sistema. Sin embargo, si la ampliación a 35 se acompañara de la elección de los 25 mejores años, entonces generaría un incremento de la pensión inicial de un 5,6% y el sistema sería más insostenible.

1. INTRODUCCIÓN

El sistema de pensiones español está inmerso en una nueva reforma, cuya primera parte se aprobó en diciembre de 2021, donde cabría destacar cuatro elementos: la revalorización de las pensiones con el IPC; el cambio en el cálculo de la pensión demorada; la modificación de la jubilación anticipada; y el Mecanismo de Equidad Actuarial que sustituye al Factor de Sostenibilidad. La segunda parte de la reforma está actualmente negociándose, en la que los temas que se van a modificar son: Régimen de Autónomos; destope de la base de cotización y de la pensión máxima; y la ampliación del periodo de cómputo de la Base Reguladora (BR).

Nos vamos a centrar en este último elemento porque, aunque parece un tema menor, puede tener importantes consecuencias sobre la sostenibilidad del sistema según cómo acabe perfilándose. En concreto, el objetivo es analizar el impacto que tendría sobre la pensión inicial y la sostenibilidad del sistema la modificación del cálculo de la BR en dos supuestos: el primero, ampliando el periodo de cómputo de los 25 años actuales a 35; el segundo, que dicha ampliación vaya acompañada de la posibilidad de elegir los 25 mejores años dentro de los últimos 35.

2. CÁLCULO DE LOS CAMBIOS EN EL PERIODO DE CÓMPUTO DE LA BASE REGULADORA

Actualmente, la base reguladora se obtiene como el promedio de las bases de cotización de los últimos 25 años antes de la jubilación, capitalizadas con la variación del IPC desde la fecha de cotización hasta 2 años antes de la jubilación. Las bases de los últimos 24 meses se incorporan al cálculo por su valor nominal. Además, en el Régimen General, se procede a la integración de lagunas de cotización, por el cual se rellenan aquellos meses donde no se ha cotizado en su totalidad. Actualmente, las 48 primeras lagunas mensuales se integran por el 100% de la Base Mínima, y las restantes por el 50% de esa misma base. Sin embargo, en el Régimen de Autónomos no se procede a la integración de lagunas.

Para realizar los cálculos se ha utilizado como punto de partida la Muestra Continua de Vidas Laborales de 2019 (en adelante, MCVL2019), calculando el impacto sobre la pensión inicial; así como la proyección futura del gasto en pensiones, tanto en términos de caja como en términos actuariales, combinando para ello datos de la MCVL2019 y de la propia Seguridad Social.

En el caso de ampliar a 35 años, como podemos ver en la tabla 1, la disminución promedio de la pensión sería del 8,7%, saliendo más perjudicados (señalados en rojo) los trabajadores con menos años cotizados, con BR elevadas, los que se acojan a la jubilación demorada, las mujeres y los del régimen de autónomos.

Sin embargo, los resultados son totalmente diferentes si la ampliación a los 35 años va acompañada de la posibilidad de seleccionar los mejores 25, ya que daría lugar a un aumento promedio de la BR de un 5,58%. Los menos favorecidos (señalados en rojo) no coinciden exactamente con los más perjudicados en el caso de la ampliación a 35 años. Ahora, saldrían menos favorecidos los trabajadores con más años cotizados, los que se jubilan anticipadamente, las mujeres y los del Régimen General frente a los autónomos. Este último colectivo tiene el inconveniente de que no se integran las lagunas de cotización y, por lo tanto, la elección de los 25 mejores años elimina muchos periodos sin cotización. Las mayores diferencias se dan en el caso de la cuantía de la BR, ya que ésta sólo aumentaría un 3,4% para los trabajadores con mayores bases, pero se incrementaría hasta el 11,82% en el caso de los de menores bases.

Tabla 1. Base reguladora promedio y variaciones con datos de la MCVL2019, según distintas características: 25 años, 35 años y 35 años eligiendo los 25 mejores. Elaboración propia a partir de Devesa et al. (2021a) (2021b).

		BR 25 años	BR 35 años	BR 35 años con 25 mejores	Variación 25 a 35 años	Variación 25 a 35 años con 25 mejores
Todas las pensiones		**1.468,27**	**1.339,87**	**1.550,16**	**-8,70%**	**5,58%**
Base reguladora en la MCVL2019	Cuartil 1	2.667,75	2.422,97	2.758,51	-9,20%	3,40%
	Cuartil 2	1.581,51	1.437,59	1.666,36	-9,10%	5,37%
	Cuartil 3	998,24	918,84	1.076,57	-8,00%	7,85%
	Cuartil 4	622,82	577,58	696,43	-7,30%	11,82%
Duración de la carrera laboral	Cuartil 1	1.822,72	1.686,42	1.905,31	-7,50%	4,53%
	Cuartil 2	1.766,45	1.638,19	1.859,33	-7,30%	5,26%
	Cuartil 3	1.491,03	1.358,30	1.588,16	-8,90%	6,51%
	Cuartil 4	790,99	674,79	846,16	-14,70%	6,97%
Edad de jubilación	Demorada (>65a 8m)	993,34	879,33	1.061,91	-11,50%	6,90%
	Ordinaria	1.356,28	1.237,34	1.436,43	-8,80%	5,91%
	Anticipada (<65a)	1.935,39	1.783,65	2.028,28	-7,80%	4,80%
Género	Hombre	1.644,54	1.516,15	1.741,08	-7,80%	5,87%
	Mujer	1.260,14	1.131,74	1.324,75	-10,20%	5,13%
Régimen	General	1.660,05	1.519,15	1.743,66	-8,50%	5,04%
	Autónomos	880,35	790,26	957,01	-10,20%	8,71%

Una opción intermedia es la de ampliar el periodo de cómputo a 35 años, pero eligiendo un número de años que pudiera permitir conseguir un determinado objetivo, por ejemplo que esta ampliación sea neutral en cuanto a la variación de la BR promedio. Para ello habría que elegir los mejores 29 años y 5 meses. En la tabla 2 podemos ver el efecto de elegir, dentro de los 35 últimos años, los mejores años entre 25 y 35. El aumento del número de años produce una disminución de la BR, con valores inferiores al valor base (últimos 25 años) para el caso de 30 o más años seleccionado.

Tabla 2. Efecto de elegir las mejores bases dentro de los últimos 35 años. Elaboración propia a partir de MCVL2019.

	Cuantía	Efecto
BR 25 años	1.468,27	
BR 35 años con mejores 25	1.550,16	5,6%
BR 35 años con mejores 26	1.532,00	4,3%
BR 35 años con mejores 27	1.513,28	3,1%
BR 35 años con mejores 28	1.494,07	1,8%
BR 35 años con mejores 29	1.474,42	0,4%
BR 35 años con mejores 30	1.454,32	-0,9%
BR 35 años con mejores 31	1.433,54	-2,4%
BR 35 años con mejores 32	1.412,28	-3,8%
BR 35 años con mejores 33	1.390,09	-5,3%
BR 35 años con mejores 34	1.366,60	-6,9%
BR 35 años con mejores 35	1.339,87	-8,7%

3. PROYECCIÓN DEL AHORRO/GASTO ANUAL GENERADO POR LA AMPLIACIÓN DEL PERÍODO DE CÓMPUTO DE LA BASE REGULADORA

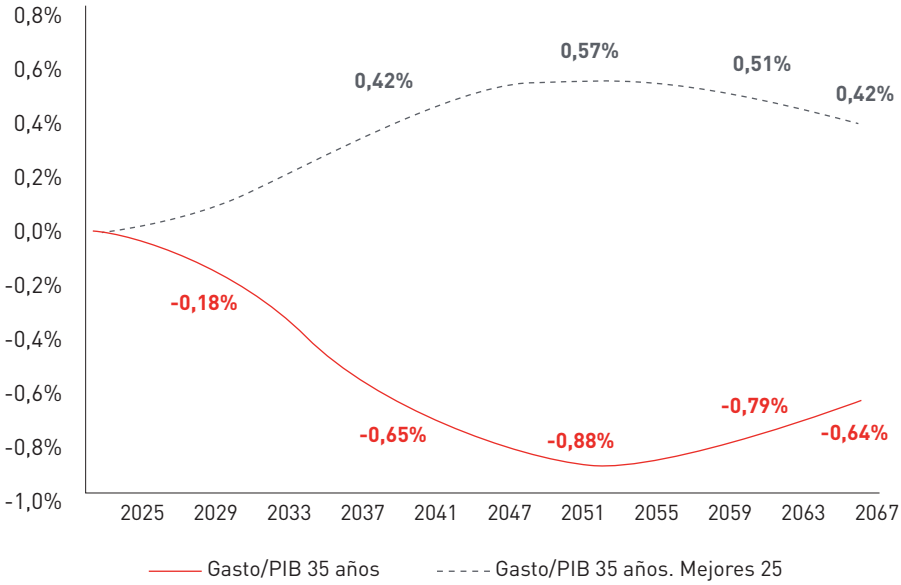
En general, todas las modificaciones que afectan solo a las nuevas pensiones de jubilación, como el caso que nos ocupa, tienen un efecto limitado a corto plazo, ya que se necesitan varios años para que lleguen a la jubilación varias cohortes, de tal forma que el efecto se vaya acumulando y llegue a ser apreciable.

Nuevamente, se va a cuantificar el ahorro anual que se generaría de ampliar a 35 años el periodo de cómputo de la BR, así como el mayor gasto en pensiones que tiene la ampliación a 35 años con la elección de los 25 mejores. Todo ello se va a realizar tanto en términos de caja como de valor actual actuarial.

Se puede ver en el gráfico 1 que, en términos de caja, en el caso de elegir los mejores 25 años se incrementa el gasto rápidamente y de forma continua (debido al periodo transitorio de 10 años que hemos supuesto) hasta estabilizarse alrededor del 0,57% del PIB durante varios años y posteriormente reducirse ese mayor gasto, llegando el último año de cálculo, 2067, a un 0,42% del PIB de ese año. En

cuanto al efecto de ampliar solo el número de años de cómputo, de los 25 actuales a 35, el efecto es el contrario, llegando a un ahorro anual máximo del 0,88% del PIB para ir disminuyendo hasta el 0,64% del PIB en 2067.

Gráfico 1. Proyección del ahorro/gasto en pensiones sobre el PIB, en términos de caja, 2023-2067 por la ampliación del periodo de cómputo de la BR. Elaboración propia a partir de Devesa et al. (2021b).

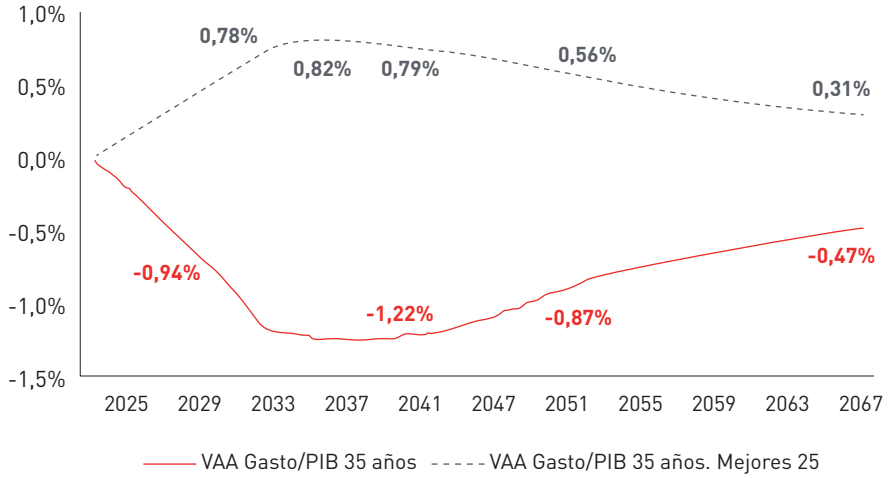


El siguiente paso es calcular el ahorro en pensiones medido a través del Valor Actual Actuarial (VAA); es decir, actualizando el mayor gasto o ahorro generado por cada cohorte, ajustado por probabilidades de supervivencia y por el factor financiero. El dato se calcula en porcentaje sobre el PIB de cada año.

Se puede ver en el gráfico 2 que, en el caso de ampliar el cálculo de la BR a 35 años con la selección de los 25 mejores, el gasto medido en VAA aumenta rápidamente hasta 2032, ya que cada año desde 2023 hasta 2032 se va ampliando un año más el período de cálculo de la BR, con lo cual el valor de la pensión inicial de las nuevas altas va aumentando un 5,58% anual, que es la referencia que se ha tomado para hacer los cálculos. A partir de 2032, donde se alcanza un mayor gasto del

0,82% del PIB de ese año, se inicia un periodo de cierta estabilidad en el que ese mayor gasto en función del PIB se mantiene en un porcentaje similar, hasta que empieza a disminuir su cuantía conforme va bajando el peso de las cohortes que accedieron a la jubilación en los primeros años. En 2067, último año de la proyección, el mayor gasto sería de 0,31% del PIB de ese año. Respecto a la ampliación a 35 años sin elección de los mejores, se observa una variación casi simétrica a la función anterior. En este caso, el menor gasto en VAA tendría un valor más elevado (en términos absolutos) de -1,22% del PIB en 2039, acabando en -0,47% del PIB en 2067.

Gráfico 2. Proyección del ahorro/gasto en pensiones/PIB, en VAA, 2023-2067, por la ampliación del periodo de cómputo de la BR.



4. CONCLUSIONES

El aumento del número de años de cómputo, de 25 a 35, para el cálculo de la BR es una medida que mejoraría la sostenibilidad y contributividad del sistema, al aumentar la relación entre aportaciones y prestaciones, pero los resultados son totalmente diferentes si, como es el caso analizado, en el cómputo de 35 años se incluyen solo los 25 mejores, porque por un lado aumenta el gasto en pensiones y

hace que el sistema sea menos sostenible, y, por otro, empeora la equidad contributiva.

Por lo tanto, la ampliación del periodo de cómputo puede ser una medida con un final muy diferente según que haya elección o no de los mejores años. Una opción que también hemos comentado es la de que tenga un efecto neutro en cuanto a gasto, aunque no en cuanto a contributividad, ya que ésta empeoraría. Para conseguir este objetivo habría que seleccionar las bases de cotización de los mejores 353 meses, es decir, 29 años y 5 meses.

BIBLIOGRAFÍA

Devesa, E.; Devesa, M.; Domínguez, I., Encinas, B. y Meneu, R. (2021a). "Efectos de la ampliación a 35 años del cálculo de la Base Reguladora en el sistema de pensiones de jubilación español". Grupo de Investigación en Pensiones y Protección Social. https://www.uv.es/pensiones/docs/pensiones-jubilacion/BR35.pdf

Devesa, E.; Devesa, M.; Domínguez, I., Encinas, B. y Meneu, R. (2021b). "Efectos de la ampliación a 35 años del cálculo de la Base Reguladora, eligiendo los 25 mejores, en el sistema de pensiones de jubilación español". Grupo de Investigación en Pensiones y Protección Social. https://www.uv.es/pensiones/docs/pensiones-jubilacion/BR35_mejores25.pdf

Muestra Continua de Vidas Laborales 2019. Ministerio de Inclusión, Seguridad Social y Migraciones.

DEVELOPMENT OF A DATAFRAME AND A BOT TO PREDICT NFT-COLLECTION PERFORMANCE

Marc Durban
Universitat Politècnica de Catalunya, España
marc.durban@gmail.com

Joaquim Gabarró
Universitat Politècnica de Catalunya, ALBCOM. CS Dept, España
gabarro@cs.upc.edu

ABSTRACT

There has been an enormous growth of blockchain technologies at the end of 2021. Particularly one of its assets called NFTs, an asset that is a new form of digital scarcity. This work aims to study their still highly volatile market and develop both a dataFrame and a bot, capable of predicting a NFT collection performance in the near future Durban (2022). We focus on the NFT main marketplace OpenSea and specifically the collections in the Ethereum ecosystem, which has the 80% of the market as of early 2022. Because of its novelty, few applications and studies have been done on this field to date. However, this has not stopped this new asset to rapidly rise in popularity from a small 10k users to more than 1.5M only in the last year. It is true that as interest grows more and more information and projects are being published, but none similar to this one (as of starting date February 2022). Despite this, other predictive applications developed for economic fields such as the stock market, might be a helpful reference.

1. WHAT ARE A BLOCKCHAIN, A NFT AND A BOT?

A *blockchain* is a list of blocks linked using cryptography Durban (2022). They are used to make decentralized transactions between users. Multiple transactions get included into a block, verified and afterwards added to the ledger. The ledger is a decentralized, shared and immutable list of verified blocks that serves as reference for everyone in the network.

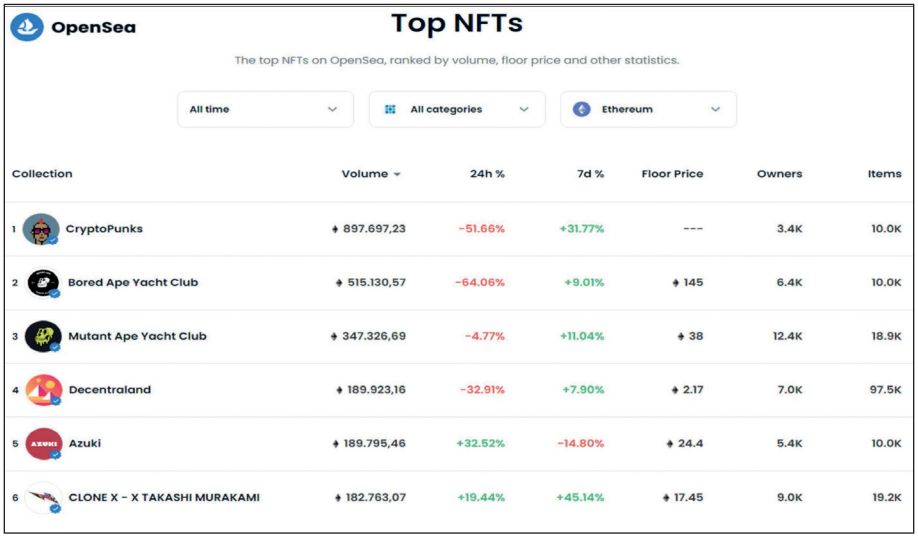
A *Non Fungible Token* or NFT, is a token/asset linked to a blockchain which is unique, indivisible and tradeable. NFTs aim to create verifiable digital scarcity and are commonly grouped, under a same brand, in what is known as *collections*. Usually they are traded in *marketplaces* such as OpenSea, looksRare, Solanart or Magic Eden.

A *bot* is a software application that runs automated tasks. Often they are used to perform tasks that are simple and repetitive.

2. ON DATA SELECTION

When trying to perform a data analysis, the size of the sample is a key factor Burkov (2019). Do not forget that data is scarce when we are talking about NFT transactions. In our case, we make use of the OpenSea ranking, (look at https://opensea.io/). In which, as we can see in Figure 1, we can consult the top best-selling NFT collections in the platform.

Figure 1. Top Ethereum collections on OpenSea, look at https://opensea.io/rankings



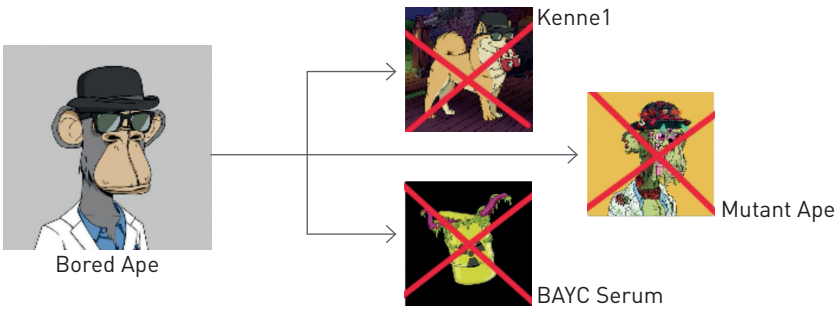
To realize this project 36 collections from this ranking have been chosen. Ensuring with this a more than enough volume of transactions for each of the collections.

Figure 2. Selected collections Durban (2022).

1	Adam Bomb Squad	13	Galactic Apes	25	My Curio Cards
2	Bored Ape Yatch Club	14	Galaxy Eggs	26	NFT worlds
3	Boss Beauties	15	Hashmasks	27	ON1 Force
4	Capsule House	16	Heart Project	28	Pudgy Penguins
5	Cool Cats	17	Jrny Club	29	Robotos
6	Creature World	18	Kaiju Kingz	30	Sneaky Vampire Syndicate
7	Crypto Toadz	19	Lazy Lions	31	Sup Ducks
8	Cyber Kongz	20	Lost Poets	32	The Doge Pound
9	Dead Fellaz	21	Meebits	33	Vee Friends
10	Doodles	22	Mekaverse	34	Wolf Game
11	Fluf World	23	Moon Cats	35	World of Women
12	Frontier Game	24	Mutant Cats	36	World Wide Web Land

Another thing to consider is the homogeneity of the dataset. Reducing the variables to the minimum, to obtain reliable results. Because of this, secondary collections from already established brands has been discarded (Figure 3).

Figure 3. Full Bored Ape Yatch Club project. Main collection plus three derived collections. All the derived collections are discarded.



Last but not least, a limited time frame has been decided to compare all dataFrames in the same timeline. Going from a starting date of December 1st, 2021 to April 30, 2022.

3. THE DATAFRAME

Once the collections of interest are selected, we start with the gathering and construction of a dataFrame, in order to deal with NFT collections. The objective of the dataFrame, from Python's data analysis library Pandas, is to store relevant information for future evaluation and analysis of a given NFT collection.

The data is extracted directly from the blockchain transactions using the API of main NFT marketplace OpenSea. Thousands of transactions are processed and key information such as the average selling price and the number of units sold on any given date are obtained.

It is important to note that there are some rare cases where transactions have been stored differently, because of a registration error or having other particularities. As less than 0.1% of the 212.729 analyzed where affected, they are ignored.

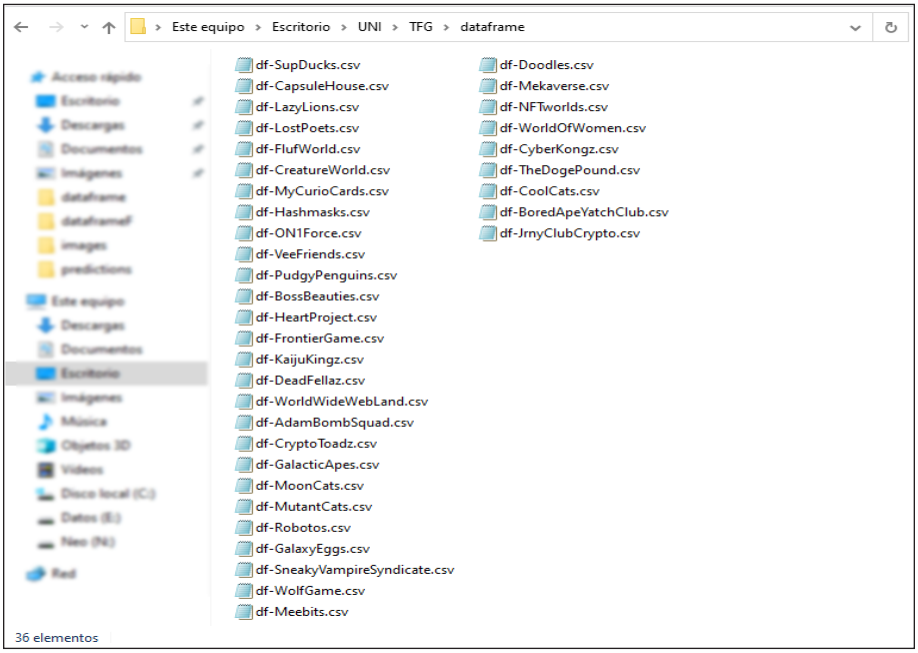
On the other hand, it has been also necessary to take care of data particularities. As one of the most important metrics for this project to track NFT collection performance has been the *average selling price on a given date*. There have been two scenarios where issues have presented for this metric. The first being when there have been 0 transactions in a given date. Making it impossible to calculate an average price and therefore setting it to 0$. The later being the case where few weird pricing sales have taken place on a day. Making the average selling price of that day to be radically different from the ones on the previous and next dates. These two cases create breaks in the continuity of the evolution of prices overtime, hurting the ability of the bot to better train and predict them. Because of this reason it has been decided to copy the average selling price of the previous day when a situation like the ones described is encountered in a dataFrame. This solution only had to be applied in 34 out of the 4356 day-prices in the 36 final dataFrames, which equals to less than 0.008% of the cases.

To the data obtained using the OpenSea API we add the average price of the Ethereum token in each date, currency with which it is exchanged in the analyzed market. This information is obtained using the API of the cryptocurrency stats provider CoinGecko.

Figure 4. DataFrame example of the Bored Ape Yatch Club NFT collection.

```
>>> pd.read_csv("./dataframe/df-BoredApeYatchClub.csv")
                PriceDoll  Sales     EthPrice
Day
2021-12-01  281055.941196     13  4637.121617
2021-12-02  243639.318996     20  4589.610618
2021-12-03  253318.435822     33  4519.441028
2021-12-04  248102.236456     31  4240.155517
2021-12-05  232848.868521     15  4101.656792
...                   ...    ...          ...
2022-03-27  356957.907103     12  3140.875711
2022-03-28  377677.026383     17  3285.173097
2022-03-29  369251.425718     11  3328.934125
2022-03-30  386368.367394     11  3401.184431
2022-03-31  427139.498417     14  3383.788762
```

Figure 5. Directory with all 36 NFT-Collection dataFrames.



As we see in the image above (Figure 4). In the case of our dataFrames, dates are used as index. Using a format of the type YYYY-MM-DD, different than the one in seconds that will be needed to carry out operations, in order to improve readability.

For any given date compressed in our limited timeframe, the average selling price of a NFT collection and its number of sales can be obtained. The first is transformed into the dollar currency, as it is a more stable measure than the Ethereum currency. This conversion is made using the average price of the Ethereum cryptocurrency on a given date and is also included into our dataFrames as the last column value.

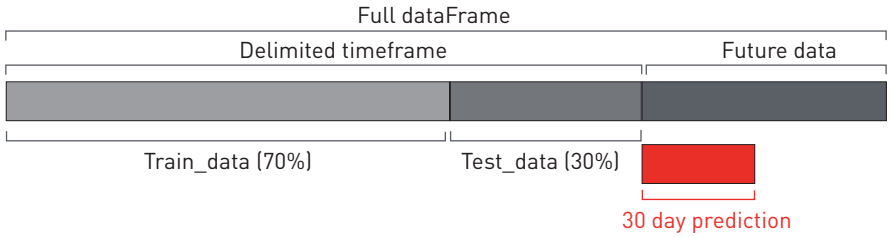
Data from every single NFT collection selected gets collected and stored in a directory called ./dataframe. A screenshot of the folder appears in Figure 5.

4. PREDICTING BOT

Remind that ./dataframe, referenced as the storage, now contains 36 .csv files. These files hold the data from each of the selected collections dataFrames, with the respective information from December 1st, 2021 to April 30, 2022.

Given a specific collection from our storage, the bot can finally start the predicting process. Information from the chosen NFT collection gets read from storage and imported into a Pandas object for better manipulation. Data then gets sliced, to obtain the amount of information we want the bot to work with. Being 4 months for the execution by default in this work. The remaining data then gets splited, using a proportion of 70 to 30, into two sets for the training and testing of the bot's model (Figure 6).

Figure 6. Dataframe partitioning for bot usage.



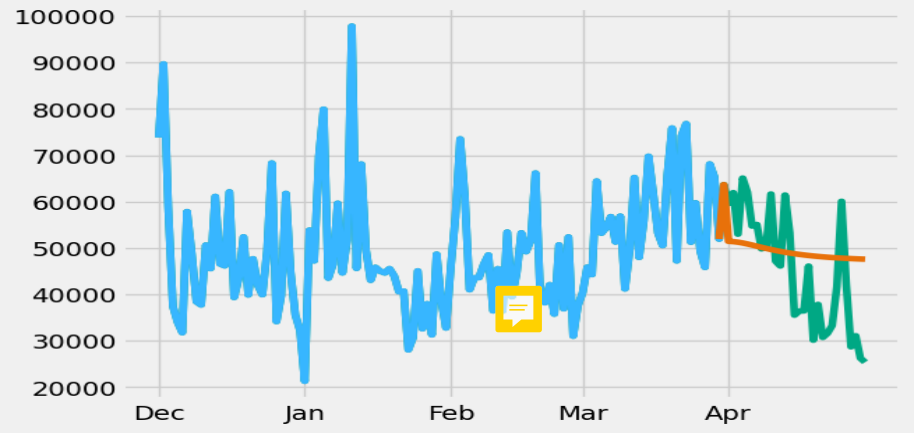
The algorithm in the bot is the LSTM (Long Short Term Memory) algorithm [Burkov (2019), Skansi (2018)]. Thanks to its ability to store key information about the past in order to make better evaluations and predictions in the future. On top of its great affinity with date-based data. This algorithm has been commonly used by many in stock predicting applications [Bhandari et al. (2022), Adusumili (2019), Sun (2020)]. This is a research domain that seems to share some similarities to the NFT market analyzed on this paper.

Using the before mentioned algorithm and the necessary training and testing data, the LSTM stacked model is then built and trained. Providing the bot with a model capable of realizing predictions for the following date, given the required data. Here is when a process named forecasting takes place. Storing the

predictions of the following date as part of the dataFrame data and this way being able to realise further predictions into the future.

An example result of the bot execution is seen in Figure 7. This execution has been made over the VeeFriends NFT collection, using 4 months of data for the model training and building. Predictions are presented by the bot using an overlapped plot with the data highlighted in different colors. In the Figure 7, colors are replaced by grey tones. First, the bot prints in blue the data used to train and build the model. In the Figure 7, this corresponds to data beginning on Dec and ending before Apr. Second, the bot highlights in green the data of our dataFrame that has not been used and will be used to compare with the prediction. In the case of the Figure 7 this corresponds to the data starting from Apr. Last but not least, the bot prints in orange the prediction. In the Figure 7, this corresponds to the more or less straight line starting at Apr.

Figure 7. Bot prediction resulting plot.



Taking in consideration that the usual data sizes for LSTM stock prediction projects, such as the ones presented in [Adusumili (2019), Sun (2020)], have at least 3 years of data. We need to take into account this fact when we consider the precision of our model's prediction for a dataset with a size of less than 6 months.

It has seemed reasonable to give in this project more weight to the predicted trend than the adjustment of these prediction to the actual values. For this reason, the average price value of both the prediction and the real data during the same period have been used for evaluation. We divide the different outcomes into three categories, this being: up-trending, U, sideways moving, S, or down-trending, D. We have considered as a successful prediction one that is in the same category as the real outcome (Figure 8).

Figure 8. Trending categories

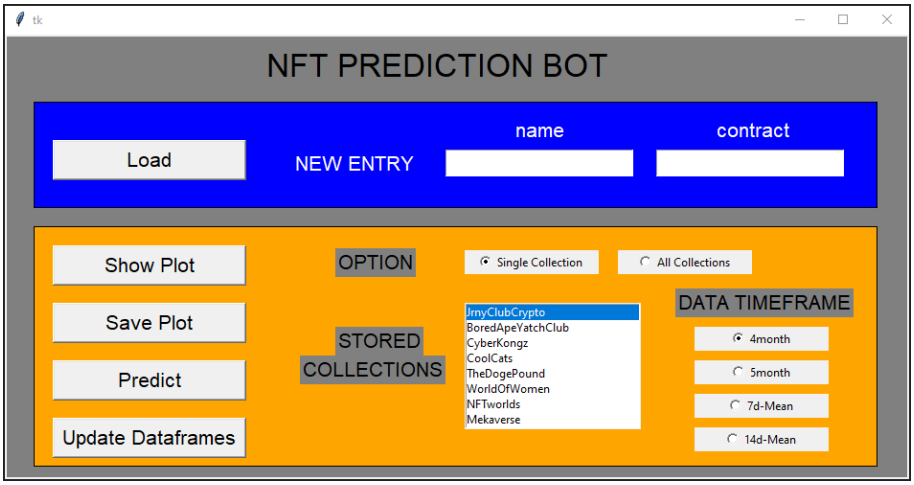
ID	NAME	OUT	PRED	ID	NAME	OUT	PRED
1	Adam Bomb Squad	S	U	19	Lazy Lions	S	S
2	Bored Ape Yatch Club	S	S	20	Lost Poets	U	U
3	Boss Beauties	S	U	21	Meebits	S	D
4	Capsule House	D	D	22	Mekaverse	S	U
5	Cool Cats	D	S	23	Moon Cats	S	U
6	Creature World	S	U	24	Mutant Cats	S	U
7	Crypto Toadz	S	S	25	My Curio Cards	S	U
8	Cyber Kongz	S	U	26	NFT worlds	D	U
9	Dead Fellaz	S	U	27	ON1 Force	S	U
10	Doodles	S	S	28	Pudgy [illegible]uins	U	S
11	Fluf World	S	U	29	[illegible]	D	U
12	Frontier Game	U	U	30	Sneaky Vampire Syndicates	S	U
13	Galactic Apes	S	U	31	Sup Ducks	D	U
14	Galaxy Eggs	D	U	32	The Doge Pound	S	U
15	Hashmasks	S	U	33	Vee Friends	D	S
16	Heart Project	D	U	34	Wolf Game	U	U
17	Jrny Club	S	U	35	World of Women	S	S
18	Kaiju Kingz	S	U	36	World Wide Web Land	S	S

As we see in Figure 9, predictions change when we pass from 4 months to 5 months.

Figure 9. Experiment result comparison between 4 months and 5 months of data used by the bot.

Finally let us consider bot's graphic interface. In order to improve data processing and automatize some actions like result extraction. A simplified interface with some key functionalities have been developed for this project to better interact with the bot and the dataFrame files. The Python library Tkinter has been used to develop this task.The minimalistic appearance of this interface is the result of prioritizing its functionality rather than its aesthetics. An image of how the mentioned layout looks could be seen in the following screenshot (Figure 10).

Figure 10. Screenshot of the graphic interface designed for the bot.



5. CONCLUSIONS

Focusing only on the Ethereum market and more specifically on OpenSea, seems to have been the right decision. For instance, a direct competitor as looksRare marketplace, which was rapidly growing at the start of the year, ended up dropping in popularity after a few months. On the contrary, OpenSea remained as the most important NFT market in the space. This together with the gained access to the OpenSea developer's API, has helped to gather all the needed data from the desired collections.

On the other hand, implementing a bot that makes predictive analysis using neural networks has been challenging. When it comes to its accuracy and optimistic predictions, found mismatches could be a product of two main factors. First, the unexpected general economic crash produced by historical facts such as the Ukraine and Russian conflict (escalating in April 18) and the increasing interest rates fruit of the covid (communicated in many instances during the 2nd Quarter of 2022). Events the model developed for this application does not have in consideration and affected all economic markets, the NFT one included. Second, the lack of more than 6 months of data for the NFT asset class. Being maybe too new to be properly analyzed with machine learning algorithms like the LSTM stacked model presented on this project.

ACKNOWLEDGMENTS

Joaquim Gabarró has been financed by: Proyecto PID2020-112581GB-C21 financed by MCIN/AEI/10.13039/501100011033.

REFERENCES

Durban, M. (2022). "Development of a dataFrame and a Bot to predict NFT-collection performance". Degree Final Project in Informatics and Engineering, Facultat de Informàtica de Catalunya, Universitat Politècnica de Catalunya, June 2022.

Burkov, A. (2019). Tthe hundred-page machine learning book". Neural Networks and Deep Learning, 61-76.

Skansi, S (2018). "Introduction to deep learning - from logical calculus to artificial intelligence". Recurrent Neural Networks, 135-152.

Bhandari, H.N., Rimal, B., Pokhrel, N.R., Rimal, R., and Khatri, R.K.C. (2022). Machine Learning with Applications, "Pedicting stock market index using LSTM". Article 100320.

Adusumili, R. (2019). "Artificial Intelligence and its Applications in Finance - Utilizing a Keras LSTM model to forecast stock trends".

Sun, L. (2020). "LSTM for stock price prediction - Technical Walk-through on LSTM-based Recurrent Neural Network Creation for Google Stock Price Prediction".

MODELLING THE IMPACT OF COVID-19 PANDEMICS ON HEALTH INSURANCE ASSOCIATED SERVICES DEMAND

Amanda Fernández-Fontelo
Departament de Matemàtiques, Universitat Autònoma de Barcelona, Spain
amanda@mat.uab.cat

Pedro Puig
Departament de Matemàtiques, Universitat Autònoma de Barcelona, Spain
ppuig@mat.uab.cat

Montserrat Guillen
Universitat de Barcelona, Department of Econometrics, Statistics and Applied Economics, RISKcenter-IREA, Spain
mguillen@ub.edu

David Moriña
Department of Econometrics, Statistics and Applied Economics, Universitat de Barcelona, Spain
dmorina@ub.edu

ABSTRACT

Health insurance is one of the branches of insurance with the greatest penetration in the Spanish market; the same happens in many developed countries. The claim rate in health insurance has suffered the impact of the Covid-19 pandemic in 2020 and 2021, especially regarding consultations and medical events that could be postponed. Mobility restrictions led to a decline in the use of insurance by policyholders and a transformation of the interaction between patients and health workers with greater use of telephone consultation. This work aims to study how to determine if, (i) due to the effect of postponing visits or (ii) due to the consequences of having suffered the virus

(persistent Covid or side effects), there will be an excess of claims and, where appropriate, when a claiming increase will occur.

1. INTRODUCTION

In the last months, there has been a big global concern around the 2019-novel coronavirus (SARS-CoV-2) infection, prompting the World Health Organization (WHO) to declare a public health emergency in early 2020. The pandemic induced by this virus has significantly impacted many aspects of human activity. In addition to the direct consequences concerning deaths caused by the Covid-19 infection and the saturation of health systems in many countries (including Spain and neighboring countries), a decrease in the use of health services has been detected in both the public health system and services associated with private health insurances in 2020.

With more than 12 million insured, health insurance is one of the insurance branches with the highest penetration in the Spanish market. More than 25% of the population has this type of coverage, exceeding 35% in some areas (UNESPA, 2020). In 2020 and 2021, the Covid-19 pandemic impacted the companies' claim rates, especially regarding medical consultations and actions that were apparently of low priority and thus were postponed. The mobility restrictions led to a decline in the use of insurance services, as well as a transformation of the interaction between patients and health workers, with a larger use of telephone consultations. The concern is whether there will be an excess of claims in 2022 and later years, either due to postponing visits or due to the pandemic consequences (e.g., because of persistent Covid symptoms or secondary effects). In the public health system, there is already evidence of an increased frequency of use of health services. However, it is not straightforward to determine if the higher frequency of claims that will be observed will be equal to or greater than the infra-loss rate that was observed during the pandemic period. For this purpose, we present here a new method for misreporting (i.e., under-reporting or over-reporting) estimation, which builds on Fernández-Fontelo et al. (2016, 2019). We use this method to determine: (i) whether the rebound effect occurs uniformly for all health insurance coverages or only for certain health insurance coverages, (ii) whether the rebound effect occurs homogeneously or varies depending on the insured characteristics, and (iii) to estimate the

time point at which the initial benefit level is recovered. Some analysis will be conducted on the effects of health spending at the public system level, but the implications for private health insurance will also be of interest. Above all, it is expected that in order to monitor the effects of the pandemic in the coming years, these forms of analysis --such as the one presented in the current work-- will be used as population groups with different sociodemographic characteristics or impacts on the use of health services cannot be directly compared.

This work aims to estimate and evaluate the pandemic's impact on health insurance, estimating the degree of under-reporting of health insurance usage, mainly in 2020, as well as the degree of over-reporting of health insurance usage nowadays as a result of the pandemic consequences. In addition, we want to develop a system that monitors claims to detect changes in the dynamics of medical insurance use in particular and any other branch in general. Although the main focus of the current research is on the development of a new method for misreporting estimation, which has also been empirically tested using simulated data, this method can also be applied to aggregated (anonymized) data from the health portfolio. We believe that the results and conclusions derived in the current research can be extended to other branches, as well as used to assess potential inequalities between countries or regions. In 2020, it was estimated that the total benefits provided by health insurance in Spain reached 6,300 million euros, of which 6,200 million correspond to the provision of medical services. In 2019, it was estimated that the total benefits provided by this type of insurance was 6,600 million euros, of which 6,500 million correspond to the provision of medical services.

The current work is organized as follows. Section 2 presents the model, its properties, and a method for estimating the model's parameters. Section 3 shows the main results of a preliminary simulation study to evaluate how the model behaves in practice in both the under-reporting and over-reporting scenarios.

2. THE MODEL

Let X_n be the following stationary INAR(1) process: $X_n = \alpha \circ X_{n-1} + Z_n$, where $\alpha \circ X_n$ is the so-called binomial thinning operator and $Z_n \sim Poisson(\lambda)$. Then, X_n follows a

Poisson distribution with $E(X_n) = \mu_X = \lambda/(1 - \alpha) = \sigma_X^2 = V(X_n)$. Suppose that the processes X_{n-1} and Z_n are independent for all n. In addition, the auto-covariance and auto-correlation functions of this process are respectively $\gamma(k) = Cov(X_n, X_{n+k}) = \alpha^k \sigma_X^2$ and $\rho(k) = Cov(X_n, X_{n+k}) = \alpha^k$.

Let Y_n be an observed and potentially misreported (i.e., under-reported or over-reported) process such that:

$$Y_n = \begin{cases} X_n \text{ with probability } 1-\omega \\ \vartheta \diamond X_n \text{ with probability } \omega \end{cases}, \quad (1)$$

where $\vartheta \diamond X_n$ is the so-called fattering-thinning operator defined as:

$$[\vartheta \diamond X_n | X_n = x_n] = \sum_{j=1}^{x_n} W_j \ (2), \quad (2),$$

and W_j are independent and identically distributed (i.i.d) random variables distributed following the probability mass function (pmf) defined below:

$$P(W_j = k | \phi_1, \phi_2) = \begin{cases} 1-\phi_1-\phi_2 \text{ if } k=0 \\ \phi_1 \text{ if } k=1 \\ \phi_2 \text{ if } k=2 \\ 0 \text{ otherwise} \end{cases}, \quad (3)$$

with $\vartheta = (\varphi_1, \varphi_2)$. Therefore, the observed process Y_n coincides with the true process X_n with probability $1 - \omega$; otherwise, the observed process Y_n is either lower (under-reported) or higher (over-reported) than the true process X_n with probability ω. The parameter $0 < \omega < 1$ is the frequency of the misreporting phenomenon, and parameters ϕ_1 and ϕ_2 indicate whether our process is under-reported or over-reported (i.e., if $\phi_1 + 2\phi_2 > 1$ the process is over-reported; if $\phi_1 + 2\phi_2 < 1$ the process is under-reported). Note that if $\phi_1 = 1$, the process is not misreported. In addition, a less flexible version of the operator defined in expression (2) can be derived if $W_j \sim Binomial(2, \phi)$. Although the distribution in expression (3) is the easiest way to model over-reporting in count time series, other probability distributions with no compact support (e.g., the Poisson or Geometric distributions) can be taken into account here as well.

The operator defined in expression (2) follows a 2nd-order Hermite distribution with parameters $\mu_X \phi_1$ and $\mu_X \phi_2$, where $\mu_X = \lambda/(1 - \alpha)$ is the expectation of X_n and

$\phi_j = P(W = j)$ for $j = 1, 2$. We can see the above using the idea of probability generating functions (pgf) as follows: First, the marginal distribution of the observed process $X_n \sim Poisson(\mu_X)$ with pgf $G_X(s) = e^{\mu X(s-1)}$. Second, the pgf of the random variable W in expression (2) is given by $G_X(s) = (1 - \phi_1 - \phi_2) + \phi_1 s + \phi_2 s^2$. Finally, the pgf of the operation defined in expression (1) is given by $G_{\vartheta \diamond Xn}(s) = G_X(G_W(s)) = exp(\mu_X \phi_1 (s - 1) + \mu_X \phi_2 (s_2 - 1))$, which is the pgf of a 2nd-order Hermite distribution with parameters $\mu_X \phi_1$ and $\mu_X \phi_2$. Hence, the expectation and variance of the random variable $\vartheta \diamond X_n$ are $E(\vartheta \diamond X_n = \mu_X (\phi_1 + 2\phi_2)$ and $V(\vartheta \diamond X_n = \mu_X (\phi_1 + 4\phi_2)$, respectively.

We already now that the latent process is distributed as $X_n \sim Poisson(\lambda/(1 - \alpha))$ and the fattening-thinning operator follows $\vartheta \diamond X_n \sim Hermite(\phi_1 \mu_X, \phi_2 \mu_X)$, where $\mu_X = \lambda/(1 - \alpha)$. In addition, we can see that the distribution of the observed process Y_n is the following mixture of a Poisson distribution and a 2nd-order Hermite distribution:

$$Y_n = \begin{cases} Poisson(\mu_X) \text{ with probability 1-}\omega \\ Hermite(\mu_X \phi_1, \mu_X \phi_2) \text{ with probability } \omega \end{cases} \quad (4)$$

with expectation and variance defined as $E(Y_n) = \mu_X (1 - \omega(1 - \phi_1 - 2\phi_2))$ and

$$V(Y_n) = \mu_X (1 - \omega(1 - \phi_1 - 4\phi_2)) + \mu^2_X \omega (1 - \omega)(1 - \phi_1 - 2\phi_2)).$$

Finally, the auto-correlation function of the observed process Y_n is defined as:

$$\gamma(k) = Cov(Y_n, Y_{(n+k)}) = \mu_X \alpha^k (1 - \omega)(1 - \phi_1 - 2\phi_2))^2,$$

which can be derived computing $E(Y_n, Y_{(n+k)})$, $E(Y_n)$ and $E(Y_{(n+k)})$. In fact, we already know that $E(Y_n) = \mu_X (1 - \omega(1 - \phi_1 - 2\phi_2))$ (and similarly for $E(Y_{(n+k)})$ given that Y_n is a stationary process). In addition, $E(Y_n, Y_{(n+k)})$ can be computed as a function of $E(X_n, X_{(n+k)}) = Cov(Y_n, Y_{(n+k)}) + E(X_n, X_{(n+k)}) = \alpha^k \sigma_X^2 + \mu_X^2$. The auto-correlation function can be thus computed as:

$$\rho(k) = Cor(Y_n, Y_{n+k}) = \frac{Cov(Y_n, Y_{n+k})}{\sqrt{V(Y_n)}\sqrt{V(Y_{n+k})}} =$$

$$= \frac{\alpha^k (1 - \omega(1 - \phi_1 - 2\phi_2))^2}{(1 - \omega(1 - \phi_1 - 4\phi_2)) + \mu_X \omega (1 - \omega)(1 - \phi_1 - 2\phi_2)^2} = \alpha^k c(\alpha, \lambda, \omega, \phi_1, \phi_2).$$

Note that $c(\alpha, \lambda, \omega, \phi_1, \phi_2)$ is constant with respect to k, and hence the ACF of the observed process Y_n is proportional (up to a multiplicative constant) to that of the latent process X_n (i.e., the ACF is geometrically decreasing at rate k given that $0 < \alpha < 1$). Note also that we take always by definition that $\rho(0) = 1$.

2.1. MODEL INFERENCES

Parameters of the model (i.e., α, λ, ω, ϕ_1 and ϕ_2) are estimated using the likelihood function, but this is not directly tractable as described by Fernández-Fontelo *et al.* (2016, 2019). According to these authors, we can compute the likelihood function here using the forward algorithm, which is based on the so-called transition and emission probabilities. In particular, the transition probabilities are given by the conditional pmf of the stationary INAR(1) process defined as:

$$P(X_n = x_n | X_{n-1} = x_{n-1}) = \sum_{j=0}^{\min(x_n, x_{n-1})} \binom{x_{n-1}}{j} \alpha^j (1-\alpha)^{x_{n-1}-j} P(Z_n = x_n - j), \quad (5)$$

where $P(Z_n = x_n - j)$ is computed in the following with the pmf of a *Poisson*(λ) distribution. In addition, the emission probabilities here are defined as given:

$$P(Y_n = y_n | X_n = x_n) = \begin{cases} 0 \text{ if } x_n < y_n/2 \\ (1-\omega) + \omega p_n \; if \; x_n = y_n \\ \omega p_n \; if \; x_n > y_n \\ \omega p_n \text{ if } x_n < y_n, x_n \geq y_n/2 \end{cases}, \quad (6)$$

and probabilities P_n are computed using the following recursive relation:

$$p_n = \frac{1}{n(1-\phi_1-\phi_2)} \left[\phi_1 \left(x_n - (n-1) \right) p_{n-1} + \phi_2 \left(2x_n - (n-2) \right) p_{n-2} \right],$$

which was computed following Baena-Mirabete and Puig (2020). Note that we consider under-reporting when $x_n > y_n$ and over-reporting when $x_n < y_n$, but $x_n \geq y_n/2$. Therefore, if there is over-reporting, our model assumes that what we see (i.e., y_n) is at most twice what actually happens (i.e., x_n).

Finally, with the transition and emission probabilities in expressions (5) and (6), the likelihood function of the observed process Y_n is recursively computed using the forward algorithm, which is essentially based on the forward probabilities. That is, the likelihood function is thus calculated as:

$$P(Y_{1:N} = y_{1:n}) = \sum_{x_N=\frac{y_N}{2}}^{\infty} \gamma_N(y_{1:N}, x_N),$$

where the forward probabilities γ_n ($y_{1:N}$,x_N) are defined as:

$$\gamma_n(y_{1:n}, x_n) = P(Y_n = y_n|X_n = x_n) \sum_{x_{n-1}=\frac{y_{n-1}}{2}}^{\infty} P(X_n = x_n|X_{n-1} = x_{n-1})\gamma_{n-1}(y_{1:n-1}, x_{n-1}) \quad (7),$$

considering $P(X_1 = x_1) \sim Poisson(\lambda/(1 - \alpha))$. Note that expression (7) is based on the transition probabilities and emission probabilities, that in our model here are defined as given in expressions (5) and (6) respectively. In practice, we use numerical optimization procedures like the *nlm* function in **R**.

3. PRELIMINARY SIMULATED RESULTS

In order to evaluate our model capabilities for under-reporting and over-reporting detection and estimation, we have first simulated two time series, one with over-reporting and another with under-reporting. Table below provides the maximum likelihood point estimates for each of the parameters for both time series, as well as the standard errors. Our model appropriately identifies whether the time series is over-reported or under-reported, and the true values of the parameters are always contained in the 90% Wald confidence intervals. Note that, although we here know which time series is over- and under-reported, our model also provides an easy mechanism for identifying which misreporting phenomenon is present in our data. In particular, if $\phi_1 + 2\phi_2 > 1$, then we very likely have over-reporting. Contrarily, if $\phi_1 + 2\phi_2 < 1$, we may have the opposite case of under-reporting. Finally, if $\phi_1 + 2\phi_2 = 1$, we do not have over-reporting or under-reporting.

We also compared several theoretical and empirical moments for both simulated time series, such as the mean, variance, and the first auto-correlation coefficients. More precisely, for the over-reported time series, we observed a mean and

variance of 7.00 and 13.65, respectively, while the corresponding theoretical values are 6.91 and 14.31. With respect to the first auto-correlation coefficients, we observed $\hat{\rho}(1)=0.192$, $\hat{\rho}(2)=0.126$, and $\hat{\rho}(3)=0.037$, while the corresponding theoretical values are $\rho(1) = 0.2183$, $\rho(2) = 0.0655$ and $\rho(3) = 0.0196$. For the under-reported time series, the empirical mean and variance were 3.34 and 7.52 compared to the theoretical ones that were 3.40 and 6.82. The first three coefficients of the empirical auto-correlation were here $\hat{\rho}(1)=0.163$, $\hat{\rho}(2)=0.101$, and $\hat{\rho}(3)=0.073$ compared to the theoretical ones that were $\rho(1) = 0.1433$, $\rho(2) = 0.0717$ and $\rho(3) = 0.0358$.

Table 1. Results of the simulation study in the over-reporting scenario.

	Over-reporting				
	α	λ	ω	ϕ_1	ϕ_2
True parameter	0.3	3.0	0.7	0.1	0.8
Point estimate	0.3578	3.2684	0.5501	0.0771	0.8072
Std. error	0.1054	0.7352	0.1155	0.0297	0.0829

Table 2. Results of the simulation study in the under-reporting scenario.

	Under-reporting				
	α	λ	ω	ϕ_1	ϕ_2
True parameter	0.5	3.0	0.7	0.2	0.1
Point estimate	0.5184	3.0586	0.7354	0.1952	0.0784
Std. error	0.1554	0.8808	0.0890	0.0511	0.0339

ACKNOWLEDGMENTS

This research is funded by Fundación MAPFRE (Becas Ignacio H. de Larramendi 2021). Amanda Fernández-Fontelo also acknowledges the Agencia Estatal de Investigación (AEI) for funding her research with the grant "Juan de la Cierva-Incorporación" (IJC2020-045188-I / AEI / 10.13039/501100011033).

REFERENCES

Baena-Mirabete, S., and Puig, P. (2020). "Computing probabilities of integer-valued random variables by recurrence relations". Statistics & Probability Letters, 161, 108719.

Fernández-Fontelo, A., Cabaña, A., Puig, P., and Moriña, D. (2016). "Under-reported data analysis with INAR-hidden Markov chains". Statistics in Medicine, 35, 26, 4875-4890.

Fernández-Fontelo, A., Cabaña, A., Joe, H., Puig, P., and Moriña D. (2019). "Untangling serially dependent underreported count data for gender-based violence". Statistics in Medicine, 38, 22, 4404-4422.

UNESPA. (2020). "Informe Estamos Seguros 2019". URL: https://www.unespa.es/main-files/uploads/2020/12/Informe-Estamos-Seguros-2019-Pagina-individual-1.pdf.

AGGREGATION OF DEPENDENT RISKS: A BRIEF SURVEY

José María Sarabia
CUNEF Universidad, Department of Quantitative Methods, Spain
josemaria.sarabia@cunef.edu

Montserrat Guillen
Universitat de Barcelona, Departament of Econometrics, RISKcenter-IREA, Spain
mguillen@ub.edu

ABSTRACT

In this paper we review some methodologies for the aggregation of dependent risks. The distribution of the sum of dependent risks is a relevant topic in actuarial sciences, risk management and in many branches of applied probability. First, we review some models of risk aggregation when we have a portfolio of dependent risks modelled with a Farlie-Gumbel-Morgenstern (FGM) copula, and some of their extensions. Because the multivariate Pareto distributions seem to be outstanding candidates to model dependent risks, we consider the problem of risk aggregation. In this case, we have closed formulas for the individual risk model and for the collective risk model assuming different primary distributions. Finally, we review the aggregation of dependent risks in mixtures of exponential distributions. The dependence structure of this model is Archimedean, and we study in detail some specific multivariate models therein. Other recent models based on copulas are also discussed.

1. INTRODUCTION

The distribution of the sum of dependent risks is a topic that has attracted the attention of researchers in risk analysis, actuarial sciences, risk management and

in many fields of theoretical and applied probability, because there are many applied situations where the most frequent and simple form of aggregation is, precisely, the sum.

In this paper we review some of the most important recent models in risk analysis. In Section 1 we review risk aggregation, in the case that the dependency structure is defined through the Farlie-Gumbel-Morgenstern copula and some of its extensions. In Section 2 we review the aggregation in the case that the risks are defined by means of the multivariate Pareto distribution. This type of dependence between risks assumes a copula of the Clayton type. Next, in Section 3 we review the aggregation of risks in the case of mixtures of exponential distributions, where the Pareto-type risk studied above is a special case. Finally, in Section 4 we provide some conclusions.

2. AGGREGATION WITH FARLIE-GUMBEL-MORGENSTERN COPULA

First, we review some models of risk aggregation when we have a portfolio of dependent risks modelled with a Farlie-Gumbel-Morgenstern copula (Cossette et al., 2013).

The main results in this case were found by Cossette et al. (2013). For this model, we have a multivariate FGM n-copula with marginal distributions the type of mixed Erlang. Then, using these previous hypotheses the distribution of the sum is again a mixed Erlang distribution, where the coefficients are defined in Cossette et al. (2013).

In addition, these authors have obtained closed-form expressions for the contribution of each risk using the TVaR and covariance rules. These findings are illustrated with some numerical examples

An interesting extension of the previous result has been proposed by Hashorva and Ratovomirija, (2015). If we consider an extension of the FGM copula of the Sarmanov-Lee type, with marginals again of the mixed Erlang type, the summation distribution remains of the mixed Erlang type, where the parameters of the aggregate distribution can be found in the work of Hashorva and Ratovomirija,

(2015). Other results on aggregation for these classes of distributions have been proposed by Vernic (2016).

3. AGGREGATION IN MULTIVARIATE PARETO DISTRIBUTIONS

In this section, we work with Pareto-type risks. This type of distribution with heavy tails has been studied from the probabilistic, inferential, and statistical modeling point of view by Arnold (2015).

In this scenario, and in order to introduce dependency in the model, the authors use the so-called common factor models.

For the construction of these models, we use the representation of the Pareto distribution as a ratio of independent exponential and Gamma distributions. In this way, we have a multivariate distribution with Pareto-type marginals and non-negative correlation between each pair of marginals.

The resulting model is the Pareto type II multivariate distribution defined by Arnold (2015), which corresponds to a Clayton-type copula. For this model, the aggregate distribution is a beta distribution of the second kind, also called beta prime distribution (see Guillén et al., 2013 and Sarabia et al., 2016).

On the other hand, the authors consider a collective risk model based on dependence, where several general properties are studied. The authors study in detail some collective models with Poisson, negative binomial and logarithmic distributions as primary distributions. It is interesting to note that in the collective Pareto–Poisson model, the probability density function is a function of the Kummer confluent hypergeometric function, and the density of the Pareto–negative binomial is a function of the Gauss hypergeometric function.

The basic proposed model has some drawbacks. All marginal Pareto distributions share the same shape parameter. An extension of this model, assuming different shape parameters for the marginal distributions, was later proposed by Guillén et al. (2019).

4. AGGREGATION IN MIXTURES OF EXPONENTIAL DISTRIBUTIONS

The Pareto distribution discussed above has an interesting property. This property indicates that a Pareto distribution can be written as a mixture of exponentials, where the mixing distribution is a classical Gamma random variable.

It is important to note that the class of distributions that are mixtures of exponential distributions covers a large class of well-known distributions, such as Pareto, Gamma, or Weibull distributions.

Making use of the mixtures of exponentials, it is possible to extend the previous Pareto model to the case of mixture of exponential distributions, to consider its multivariate extension and, subsequently, to obtain simple analytical formulas for the aggregate distribution.

In this context, a multivariate distribution can be defined so that its marginal distributions are mixtures of exponentials (Whitmore and Lee, 1991).

This distribution has several interesting properties. For example, many of the formulas can be easily expressed in terms of the Laplace transform of the mixing distribution. On the other hand, the model can be characterized by having marginals that can be interpreted as corresponding to claim severities that are completely monotone with a dependence structure due to an Archimedean copula with generator given by the inverse of the Laplace distribution of the mixing distribution (Oakes, 1989; Albrecher et al., 2011).

The result that allows us to obtain the aggregation of dependent risks in mixtures of exponential distributions was obtained by Sarabia et al (2018). The dependence structure of this model is Archimedean. Using this result, it is possible to consider specific multivariate models with severities of the type of Pareto, Gamma, Weibull, inverse Gaussian mixture of exponentials and other parent distributions. For this model several additional properties can be obtained including the VaR and TVaR risk measures.

Some extensions of the basic multivariate model have been proposed in the literature. A similar result for the case of Bernstein copulas has recently been proposed by Marri and Moutanabbir, (2022).

ACKNOWLEDGMENTS

This work was funded (JMS and MG.) by grant no. PID2019-105986GB-C22 and ... by MCIN/AEI/10.13039/501100011033

REFERENCES

Albercher, H., Constantinescu, C., and Loisel, S. (2011). "Explicit ruin formulas for models with dependence among risks". Insurance: Mathematics and Economics, 48, 265-270.

Arnold, B.C. (2015). Pareto Distributions. Second Edition. CRC Press and Chaman & Hall Book.

Cossette, H., Cote, M.P., Marceau, E., and Moutanabbir, K. (2013) "Multivariate distribution defined with Farlie-Gumbel-Morgenstern copula and mixed Erlang marginals: Aggregation and capital allocation". Insurance: Mathematics and Economics, 52, 560-572.

Guillen, M., Sarabia, J.M., and Prieto, F. (2013). "Simple risk measure calculations for sums of positive random variables". Insurance: Mathematics and Economics, 53, 1, 273-280.

Guillen, M., Sarabia, J.M., Prieto, F., and Jordá, V. (2019). "Aggregation of dependent risks with heavy-tail distributions". International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 27, Supp01, 77-88.

Hashorva, E., and Ratovomirija, G. (2015). "On Sarmanov mixed Erlang risks in insurance applications". ASTIN Bulletin, 45, 175-205.

Marri, F., and Moutanabbir, K. (2022). "Risk aggregation and capital allocation using a new generalized Archimedean copula". Insurance: Mathematics and Economics, 102, 75-90.

Oakes, D. (1989). "Bivariate survival models induced by frailties". Journal of the American Statistical Association, 84, 487-493.

Sarabia, J.M., Gómez-Déniz, E., Prieto, F., and Jordá, V. (2016). "Risk aggregation in multivariate dependent Pareto distributions". Insurance: Mathematics and Economics, 71, 154-163.

Sarabia, J.M., Gómez-Déniz, E., Prieto, F., and Jordá, V. (2018). "Aggregation of dependent risks in mixtures of exponential distributions and extensions". ASTIN Bulletin: The Journal of the IAA, 48, 3, 1079-1107.

Vernic, R. (2016). "On the distribution of a sum of Sarmanov distributed random variables". Journal of Theoretical Probability, 29, 118-142.

Whitmore, G.A., and Lee, M.L.T. (1991). "A multivariate survival distribution generated by an inverse Gaussian mixture of exponentials". Technometrics, 33, 39-50.

MODELING AND FORECASTING TIME SERIES OF CO2 EMISSION PRICES ON THE EUROPEAN UNION EMISSIONS TRADING SYSTEM

Zbigniew Krysiak
Warsaw School of Economics, Institute of Corporate Finance and Investments, Poland
zkrysiak@sgh.waw.pl

Urszula Krysiak
Ignacy Moscicki University of Applied Sciences in Ciechanow, Institute of Finance and Investments, Poland
Urszula.krysiak@gmail.com

Krzysztof Malczewski
University of Life Sciences, Institute of Technology, Poland
krzysztof_malczewski@sggw.edu.pl

Adrian Markiewicz
University of Social Sciences, Faculty of Management and Security Studies, Poland
adrian_markiewicz@wp.pl

ABSTRACT

European Union Emissions Trading System (EU ETS), although is operating since many years, shows still very unstable and unjustified trends of CO2 prices, and it seems to be not creating the efficient market. Against falling CO2 emission in EU by about 30% over last 15 years, the prices of CO2 are still increasing, what demotivate producers of CO2 by investing in the low emission carbon technologies, since increasing costs of CO2 decline their profits creating the shortage in the capital for the new investments. In such circumstances arises the question, what models, tools, and measures may be adequate to anticipate the CO2 prices in the process of the investment projects appraisals.

In the hereby presented paper we will display the approaches, models, tools, and measures, which help to model CO2 prices, and as well identify problems to be solved for improving the quality and increasing the profitability of the low-carbon emission investment projects.

1. INTRODUCTION

CO2 prices are most important input-data to the models for the investment projects appraisal, when applying the low-carbon emission technology. There may be used different models for the project appraisals (Tomas, I., & Višić, J., 2020), (Mukhtar, W., & Agarwal, R. K., 2009), (Krysiak Z., 2021), but we are going to highlight the added value of the Binominal Model (BM), called as well as an option model.

For many energy producers cost of the CO2 allowance stands for almost 60% of total production costs, what in the consequence very much charge the consumers in form of final energy price. The research problem finally aims to support investors with collecting of the appropriate tools, models, and measures for project appraisal, so that to obtain most accurate answer if the savings due to implementation of the low emission CO2 technology counterweights the increasing price of the CO2 certificates. From that perspective it is important to make sure that implementation of project increase the value of the producer of CO2 (Mo, J-L., Zhu, L., Fan, Y., 2012), (Krysiak, Z., 2015).

On the one hand, high prices of the certificates should motivate companies operating in the energy industry for the investments leading to declining the quantity of pollution due to CO2 emission, but on the other hand too high prices of certificates will increase cost of energy and decline the profit margins, what reduce the company's potential for the new investments and reduces the value. From that perspective it is important to find the golden principle to balance between these two perspectives.

The undertaken research problem relates to having in mind the equilibrium-point between the level of the CO2 price and the cost of energy production, so-as to maximally stimulate low-carbon emission investments. The energy transformation requires huge investment outlays, which is crucial in the process of reducing

CO2 emissions, therefore, first-of-all, it is necessary to increase the capital potential of economic entities, which will enable investments to produce "clean energy".

There are many opinions that the prices of CO2 on the EU ETS do not reflect adequate cost of the pollution due to the emission of CO2. One of the biggest, currently discussed, reasons of this discrepancy are the financial institutions, who although are not producers of CO2 develop the speculative trading on the prices of CO2 certificates. Their speculative approach is focused on the profit and not on the return on investment into real low-carbon emission projects. This implies that financial institutions are not able to evaluate real investment project so that to estimate its profitability against CO2 prices, cost of certificates for CO2 emission, and permitted level of free of charge CO2 emission.

Accepted profitability of the investment project may be obtained at certain trend of CO2 prices in short and long term in correlation with the allowed level of free of charge CO2 emission. Regulators should be aware that to low permitted level of free of charge CO2 emission and to fast growth of CO2 prices will demotivate investors in real projects and will not help to speed up the decline of global CO2 emission. This context implies how to shape and reform EU ETS since current development of the CO2 prices create a big barrier for further speed of decrease of CO2 emission. From that perspective modeling the CO2 prices should not be based on historical time series for the purpose of forecasting CO2 in short term. There are many different approaches based on the forecasting time series (Song, Y. et al., 2019), (Zhu, B., Ye, S., Wang, P., He, K., Zhang, T., and Wei, Y., 2018), (Sun, W., and Zhang, C., 2018), which apply very complicated econometrics method, which we think contribute not much for the profitable real investment decisions.

Hence, it is an important issue, how to process, elaborate, and forecast the data obtained from UE ETS for their most efficient utilization by the investment project appraisal, so that to help investors to answer the question if the implementation of the specific low emission CO2 technology counterweights the increasing price of the CO2 certificates. In that context of this research problem, we put following questions:

- Does the Environment, Social, Government (ESG) approach to the business development increase the return on capital and decline the risk?

- How much the volatility assessment of CO2 prices based on the GARCH model is helpful for the low-carbon emission technology projects appraisal?

- What may be the role of the Real Volatility Distribution (RVD) of CO2 prices at Monte Carlo Simulation (MCS) in the project appraisal?

- Would be the Binominal Model (BM) efficient tool for appraising the low-carbon emission technology projects?

2. ESG APPROACH INCREASE THE RETURN ON CAPITAL TO THE BUSINESS

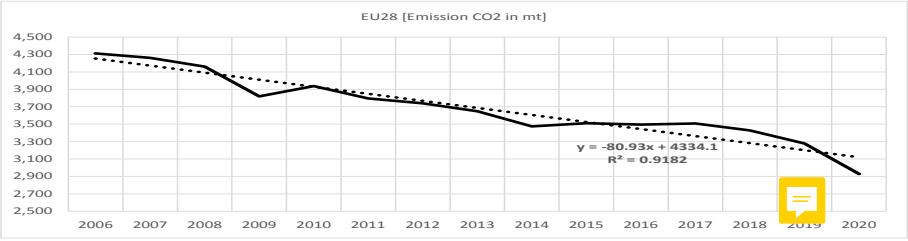
Implementation of the low-carbon emission technology projects declines CO2 emission, but for the CO2 producers most important is the positive financial impact on the profit and loss statement and return on capital. The convincing prove for the investors may be higher return on capital in the group of companies operating in the ESG concept than for the entities no focused on environment pollution. The following research-study, we made, delivers positive evidence. We analyzed 50 ESG companies and indices from all over the word against 25 main stock exchanges indices. This study proved that ESG delivers higher return on capital than stock indices by about 15.1%, which was statistically significant. The risk measure between the ESG and stock indices were not statistically different. The results of this research were present in Figure 1. Based on our studies, the return on capital for the investments aimed to protect the environment against damages resulted by carbon emission is very high and exceeds 24%.

Figure 1. Rate of return and risk on ESG and stock indices

Measure	Average of ESG	Average of Stock Exchange Indices	Difference between ESG and Stock Indices
Annual volatility as measure of risk (StdDev [σ])	22.8%	19.9%	2.9%
Annual Rate of Return [R]	24.3%	9.2%	15.1%
Sharpe'a Ratio	1.06	0.44	0.62

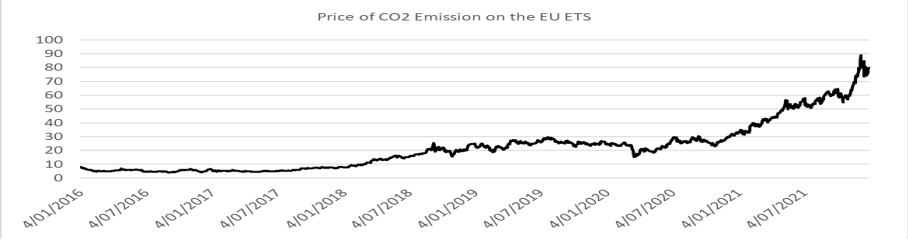
This justifies very big decline of CO2 emission in EU, which in the last 15 years dropped by about 30%, what was reflected on in the Figure 2. On the beginning of the 2006 the CO2 emission has reached the level of 4334 million ton (Mt), and afterwards was observed steady decline with the average annual rate of around 80.9 Mt., and finally at the 2022 reached the level of about 3100 Mt. This big decline of CO2 emission was due to relatively low CO2 prices on EU EST, which were in the range between 10-20 Euro as shown in the Figure 3, and because of high rate of return on invested capital into the low-carbon emission projects amounted to 24%.

Figure 2. Emission of CO2 in EU for the period 2006-2020



The rate of decline in the CO2 emission shown in the Figure 2 enables to approach zero CO2 emission target after 80 years. So, optimism in EU, who is aiming to approach zero emission goal in 2050 seems to be not realistic. After 2020, we observe strong slope up in the CO2 prices on EU EST, which creates serious barrier for next investments of the low-carbon emission projects, because of high charges with costs of the CO2 producers resulted from very high CO2 prices on EU EST. The extremally rapid growth of the CO2 prices from 2020 brought its level three times up to the 90 Euro, what was shown in the Figure 3.

Figure 3. CO2 prices on EU EST for the period 2016-2022



3. ASSESSMENT OF THE VOLATILITY IN THE GARCH MODEL

We performed studies of the volatility CO2 prices on the stock exchange EEX-Leipzig based on the GARCH (1,1) model, which showed a statistically significant GARCH effect with confidence level of 95%, whereas p-Value was less than 5%. Since the expected value of the CO2 price in the future depends on the volatility like it was shown in formula 1, model GARCH, which reflects formula 2, helps to identify the path of the historical volatility and as well the short term forecasted volatility of the CO2 prices as shown in formula 3.

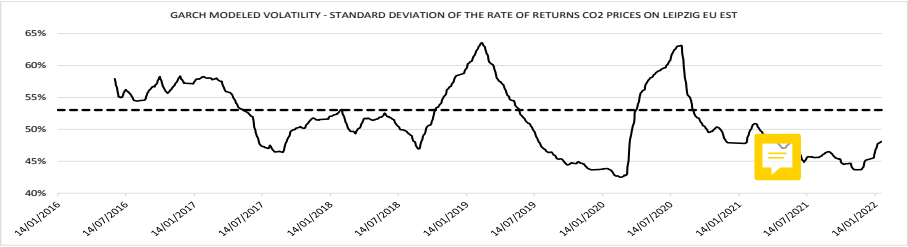
$$S_t = S_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t} \tag{1}$$

$$\sigma_n^2 = \gamma V_L + \alpha u_{n-1}^2 + \beta \sigma_{n-1}^2 \tag{2}$$

$$E[\sigma^2{}_{t+n}] = V_L + (\alpha + \beta)^n (\sigma^2{}_t - V_L) \tag{3}$$

The volatility analysis of the CO2 prices on the stock exchange EEX-Leipzig shown in the Figure 4 reflects fluctuation around the long run volatility of 53% with highest deviations up to 65% and lowest deviation up to 42%. In this research parameters of the GARCH model were equal to γ= 0.0924; α=0.1146; β=0.7930, what is associated with very high fluctuations and long time to revert to the long-run volatility on average around 60 days.

Figure 4. GARCH modeled volatility of the CO2 prices on stock exchange EEX-Leipzig in period 2016-2022

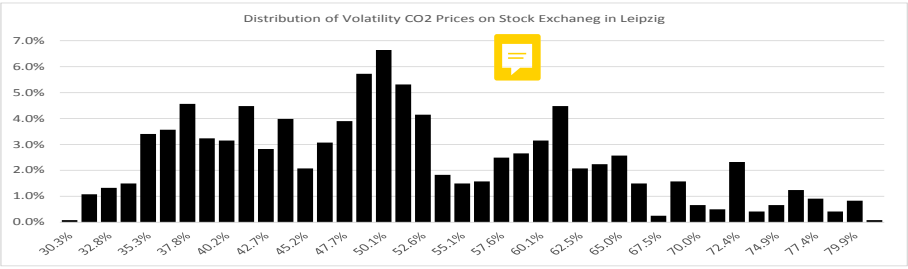


The volatility analysis applying the GARCH model is useful tool for pricing the futures contracts on CO2 prices (Byun, S.J., Cho, H., 2013), what belongs to entire and the complex process of forecasting and modeling CO2 prices.

4. REAL VOLATILITY DISTRIBUTION FOR MONTE CARLO SIMULATION IN THE PROJECT APPRAISAL

Project appraisal based on the Binominal Model needs to run Monte Carlo (MC) simulation with input data in the form real volatility distribution of CO2 prices from Leipzig stock exchange. For this purpose, was created distribution, as shown in Figure 5, what afterwards is base for creating the generator of the volatility CO2 prices, when MC simulation is processed. This enables to increase the quality of project appraisal (Krysiak Z., 2021).

Figure 5. Distribution of volatility CO2 prices on Stock Exchange in Leipzig

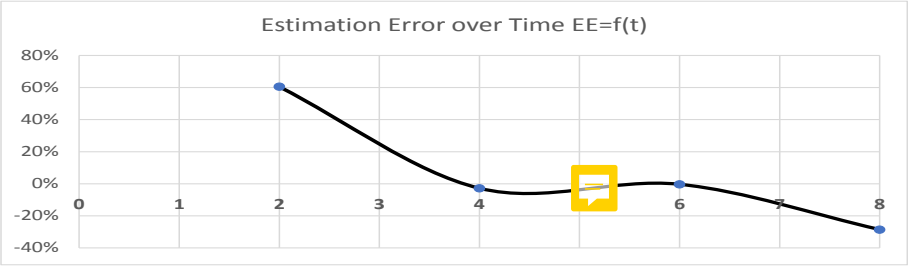


5. BINOMINAL MODEL FOR APPRAISING THE LOW-CARBON EMISSION TECHNOLOGY PROJECTS

There are positive results based on the research studies performed on the sample of the ten companies from Warsaw Stock Exchange, which aimed to test the quality of the Binominal Model (BM). The BM supported with the Monte Carlo simulation based on the real volatility distribution delivered high accuracy in the medium horizon from 4 to 6 years. Average accuracy was observed for 3 and 7 years, low accuracy for 8 years and very low accuracy for the short horizon from 1 to 2 years. In 80% of cases, the equity value was overestimated by + 23.4% or underestimated by -16.2%. An accurate estimation with the BM, was observed in 20% of cases,

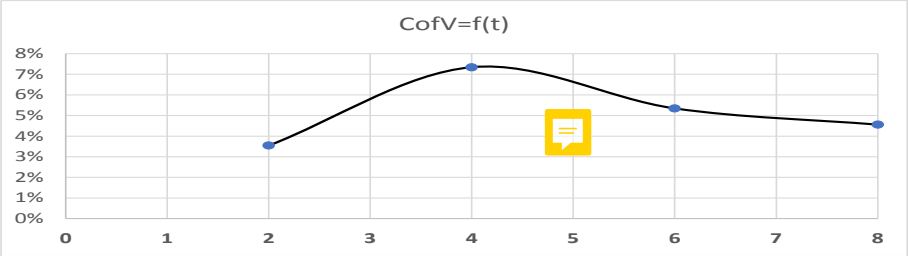
which means that the market value of equity ties in with those forecasted by the BM. On average the estimation error was + 7.2%, which seems to be very attractive for practical applications, since a forecast deviating by 7.2% from the market value provides investors and management with very a satisfactory tool supporting the decision-making processes. Figure 6 presents the estimation error in respect of time (*EE* = *f*(*t*)), where EE stands for estimation error. From this figure, the estimation errors in time horizon of 4 to 6 years are close to zero.

Figure 6. The estimation error in the BM in respect of time EE=f(t) (the vertical axis shows the estimation error, and the horizontal axis - time horizon as the subsequent years in the BM)



In Figure 7 the coefficient of variation in respect of time CofV=f(t) was presented for distribution of equity in the BM. The coefficient of variation was defined as the relation between the standard deviation of the forecasted equity to the average value of the forecasted equity in the sample under consideration (in other words, it is a quotient of the standard deviation of the forecasted equity to the average forecasted equity in the sample under consideration). The coefficient of variation indicates that the estimated value of equity deviated from the mean. On average this coefficient equaled 6.9%, which is very low if one realizes that the variation of the forecasted value of equity in a long horizon deviates only not as much.

Figure 7. The variation coefficient in the BM in respect of time EE=f(t) (the vertical axis shows the estimation error, and the horizontal axis - time horizon as the subsequent years in the BM)



The positive results indicate the high quality of the BM, but this was very much because the Monte Carlo simulation was applied with real distributions volatility supported with volatility's distribution generators built in the Excel model. Building the generators is in fact a time-consuming process but it is worthwhile indeed.

4. CONCLUSIONS

The main message we wanted to deliver in this research is that the modeling of CO2 prices for the purpose of the profitable low-carbon emission project appraisal is not the same as the CO2 prices forecast. In this context, the modeling of CO2 prices means, that we need to draw the future trend or path of CO2 prices in correlation with the variable level of permits over the time in horizon of the investment project execution. There are four components of the frame helping to deliver with high likelihood the positive financial outcome and decline in CO2 emission in the process of the investment project. In this context each time the investment project is to be appraised following four components should be considered and utilize:

- checking the current historical rates of return on capital for similar investments in the certain industry as an adequate reference,

- applying GARCH model to show the historical and future volatility characteristics,

- create the Real Volatility Distribution of CO2 prices as an input fo Monte Carlo Simulation in the process of project appraisal,

- applying and verifying the profitability of the project based on the Binominal Model.

REFERENCES

Byun, S.J., and Cho, H. (2013). "Forecasting carbon futures volatility using GARCH models with energy volatilities". Energy Economics., 40, 207-221.

Chevallier, J., Pen, Y.L., and Sévi, B. (2011). "Options introduction and volatility in the EU ETS". Resource and Energy Economics, 33, 4, November, 855-880.

Chevallier, J. (2011). "Nonparametric modeling of carbon prices". Energy Economics, 33, 6, 1267-1282.

Gibson, R., Krueger, P., and Schmidt, P.S. (2019). "ESG Rating Disagreement and Stock Returns". (December 22, 2019). Swiss Finance Institute Research Paper No. 19-67, European Corporate Governance Institute – Finance Working Paper No. 651/2020, Available on SSRN: https://ssrn.com/abstract=3433728 or http://dx.doi.org/10.2139/ssrn.3433728

Krysiak, Z. (2015). "Financial Engineering in project development". Warsaw: Warsaw School of Economics.

Krysiak Z. (2021). "Accuracy of the equity's forecast in option model simulated by real volatility distribution". Financial Sciences. Nauki o Finansach, 26, 1.

Mo, J-L., Zhu, L., and Fan, Y. (2012). "The impact of the EU ETS on the corporate value of European electricity corporations". Energy, 45, 1, September, 3-11.

Mukhtar, W., and Agarwal, R.K. (2009). "DCF valuation of a firm: A case for application of Monte Carlo simulation". Available on SSRN: https://ssrn.com/abstract=1501589 or http://dx.doi.org/10.2139/ssrn.1501589.

Shi-JieDeng, L., and Thomas, V.M. (2016). "Carbon emission permit price volatility reduction through financial options". Energy Economics, 53, January, 248-260.

Song, Y., Liu, T., Liang, D., Li, Y., and Song, X. (2019). "A fuzzy stochastic model for carbon price prediction under the effect of demand-related policy in China's carbon market". Ecological Economics, 157, 253-265.

Stefan, M., and Wellenreuther, C., (2020). "London vs. Leipzig: Price discovery of carbon futures during Phase III of the ETS", Economics Letters, 188,108990.

Sun, W., and Zhang, C. (2018). "Analysis and forecasting of the carbon price using multi-resolution singular value decomposition and extreme learning machine optimized by adaptive whale optimization algorithm". Applied Energy, 231, 1354-1371.

Tomas, I., and Višić, J. (2020). "Real option analysis – decision making in a volatile environment". Retrieved from https://bib.irb.hr/datoteka/430316.

Parry, I., Veung, C., and Heine, D. (2015). "How much carbon pricing is in countries' own interests? the critical role of co-benefits". Climate Change Economics, 6, 4, 1550019.

Zhu, B., Ye, S., Wang, P., He, K., Zhang, T., and Wei, Y. (2018). "A novel multiscale nonlinear ensemble leaning paradigm for carbon price forecasting". Energy Economics, 70, 143-157.

ANALYSIS OF CONSECUTIVE FINANCIAL STATEMENTS CONCERNING BANKRUPTCY PREDICTION

Tobias Nießner
University of Goettingen, Chair of Application Systems and E-Business, Germany
tobias.niessner@uni-goettingen.de

Matthias Schumann
University of Goettingen, Chair of Application Systems and E-Business, Germany
mschuma1@uni-goettingen.de

ABSTRACT

While classical bankruptcy prediction models focus on the analysis of financial ratios extracted from a balance sheet, there is a consensus in research that harnessing textual data from annual reports can optimize these models. Motivated by the expectations of researchers and analysts, we addressed the informative value of unstructured data by looking closely at period of a company's evolving bankruptcy over time. To this end, we used term frequency-inverse document frequency (TF-IDF) to analyze the cosine similarity of consecutive financial reports of 23 solvent and insolvent German companies over 5 years. Our results suggest that optimization of bankruptcy prediction models should not only depend on extracted key figures of individual annual reports text but should also consider the development of those within past years to arrive at a more accurate result. The changes are more revealing than the commonalities, especially concerning the analysis of textual data.

1. INTRODUCTION

While various studies deal with the prediction of corporate insolvencies, it is easy to see that the analysis of textual components of annual financial statements often focuses on static characteristics, e.g. sentiment (Myšková and Hájek, 2020) or readability (Le Maux and Smaili, 2021), of individual documents. In practice, however, an analyst may also be interested in the extent to which planned investments, but also the presentation of risks for a company, are depicted beyond these textual key figures (Lohmann and Ohliger, 2020). In this context, research findings show that the communication of companies to stakeholders can be subject to a wide variety of motives that condition the presentation of developments (Bloomfield, 2002).

In the following, the question arises to what extent an impending bankruptcy of a company can also be anticipated by stakeholders through an analysis of the development of textual components of the annual financial statements. Of particular interest is whether new information can be extracted from looking at textual data compared to previous years.

> *RQ: How does the information sharing within financial statements change about developing bankruptcy?*

We start with a presentation of our data acquisition and selection process. Then, we describe our methodology, present our findings and discuss them concerning our research question. In the end, we unfold limitations and complete our paper with a conclusion.

2. DATA COLLECTION

We refer in this study to a data set of published annual financial statements by the german central platform *Bundesanzeiger* for official announcements as well as for legally relevant company news. Therefore, within the company selection process, we identified bankrupt medium-sized companies in Germany using the *Amadeus database of Bureau van Dijk*. Based on the criterion of company size, we were able to ensure that the companies we selected are required by German law to publish

a management report within their annual financial statements (HGB, 2021a, 2021b) and thus provide a suitable amount of textual data for the company's self-presentation to external stakeholders. Furthermore, it was determined that only bankrupt companies that published their last annual financial statements after 2017 were considered. However, to avoid any bias due to temporal selection, the financial statements of the solvent peer companies from the same years were chosen. To be able to compare a suitable and comparable set of consecutive annual financial statements, it was further ensured by the *Bundesanzeiger* platform that five consecutive annual financial statements are available for the consideration of the change until the occurrence of a company's bankruptcy. To additionally ensure that the information in the database is up to date, we checked with the German portal for insolvency announcements (InsBekV, 2021) and whether official bankruptcy applications had been filed. To obtain a corresponding control group for comparison purposes, the *Amadeus database* was used to identify solvent companies in terms of company size and industry affiliation. A total of 23 solvent and insolvent companies were identified for the analysis (see Table 1).

Table 1. Overview of companies

Bankrupt companies	Solvent companies
Company name	Company Name
Alma-Küchen GmbH & Co. KG	3B GmbH
AWG GmbH	Aerologic GmbH
Bachtrup GmbH	Ascent AG
Böhm AG	BCD Travel GmbH
Curasan AG	Beckermann Küchen GmbH
Clean Garant GmbH	Christophorus Trägergesellschaft mbH
Deutsche R + S GmbH	DO & CO Holding GmbH
Envirotherm GmbH	Dr. Rehfeld Fashion AG
Esprit Retail B.V & Co. KG	Gamestop Deutschland GmbH
Galeria Karstadt Kaufhof GmbH & Co. KG	Hilfiger Stores GbmH
Germania Fluggesellschaft mbH	Holmer Maschinenbau GmbH
GerryWeber AG	IBR Gebäudem. GmbH & Co. KG
Greensill GmbH	Kaco GmbH & Co. KG
Hallhuber GmbH	Kaufland Stiftung & Co. KG
Katharina Kasper ViaSalus GmbH	Logaer Maschinenbau GmbH
Kath. Klinikum Oberhausen GmbH	NKD Deutschland GmbH

Bankrupt companies	Solvent companies
Company name	Company Name
Klier Holding GmbH	RENAFAN Holding GmbH
Thomas Cook GmbH	Stadtwerke Groß-Gerau GmbH
Spiele Max GmbH	TAKKO Holding GmbH
Vapiano SE	Vinzenz Murr GmbH
Veritas AG	Webac Holding AG
Vidrea Deutschland GmbH	Wortmann GmbH
Wilke Waldecker GmbH & Co. KG	Zara Deutschland B.V. & Co. KG

3. METHODOLOGY

To analyze the linguistic change of texts, various levels of granularity, i.e. character, word, sentence, and document, can be distinguished according to the taxonomy of Fromm et al. (2019). In this study, we consider a syntactic level of the linguistic analysis of financial statements and subsequently use the term frequency-inverse document frequency (TF-IDF) to obtain a representation of the individual documents in the form of a vector. To assess the change of the individual financial statements, which are considered in time series of five years each, we use the cosine similarity of the vectors calculated in this way among each other as a metric of change. This methodology is widely used in IS research (Bankamp et al., 2021) and has achieved good results in uncovering new and reducing redundant information (Zhang et al., 2002). Nevertheless, we are aware that there is also a strong focus in recent research on other document representation methodologies, e.g., latent dirichlet allocation, word embeddings, and document embeddings, in finance and accounting (Bankamp and Muntermann, 2022). However, since we have to start from the classical approach of documents that contain relatively rigid content over years, we consider our chosen approach as experimental in the context of this study to fundamentally show the tendency of change concerning occurring insolvency.

The cosine similarity describes the cosine of the angle, i.e. α, between two vectors, i.e. x and y, and can thus be represented by the scalar product of both vectors and the euclic norm for the distance in the following formula:

$$\textit{Cosine Similarity}(x,y) = \cos(\alpha) = \frac{x \cdot y}{||x|| \cdot ||y||} \quad (1)$$

Since we consider frequency vectors we get a range of values from 0 to 1. It is directly obvious from the formula that a smaller angle, i.e. a greater commonality of both documents, is associated with a higher value for the cosine similarity. For the individual calculations of cosine similarity and the term-frequency inverse document frequency, it is recommended to use the scikit-learn Python package, for example (scitkit-learn, 2022).

4. RESULTS AND DISCUSSION

It is directly interesting to see that just before insolvency occurs, stronger changes are evident than in solvent companies and a one-time larger difference in Y-2 to Y is also visible, while the observation of the solvent companies suggests that a constant change is present.

In the following, we use *Y* to denote the year in which an insolvency petition was filed and, consequently, *Y-n* to denote the *n*-th year before that. Since distance matrices are symmetric by definition, Table 2 and Table 3 are each upper triangular matrices. We see directly that the cosine similarities among financial statements regardless of whether bankrupt or solvent firms are considered have a very high lower bound of about 0.94.

Table 2. Distance matrix of bankrupt companies

Cosine Similarity				
Y	Y-1	Y-2	Y-3	Y-4
1	0.9688	0.9445	0.9409	0.9327
	1	0.9648	0.9561	0.9453
		1	0.9729	0.9563
			1	0.9710
				1

Table 3. Distance matrix of solvent companies

Cosine Similarity				
Y	Y-1	Y-2	Y-3	Y-4
1	0.9788	0.9631	0.9500	0.9430
	1	0.9750	0.9570	0.9484
		1	0.9723	0.9570
			1	0.9749
				1

If we also look at the upper secondary diagonal of both matrices, it also becomes clear that this difference essentially becomes more apparent when we look at a longer time series, since according to our calculations, successive annual financial statements remain more or less constant for both bankrupt and solvent companies. Nonetheless, the similarity score of consecutive financial statements for bankrupt firms averages 0.969, slightly lower than that of solvent ones at 0.975.

Our study is, of course, subject to limitations. It should be noted that Lang and Stice-Lawrence (2015) were able to show in a study using an analysis of the cosine similarity of pairwise consecutive annual financial statements of various companies from different countries that the adaptation of IFRS led to a significant increase in new information. An interesting consideration is the impact of market-changing situations, such as the COVID-19 pandemic. In *Delivery Hero's* annual financial statement, it is clear that this situation had a major impact on reporting, i.e. 65 mentions in 2020 (Delivery Hero, 2020) and 56 mentions in 2021 (Delivery Hero, 2021). It must therefore be assumed that a change in reporting can occur for a variety of reasons and cannot be regarded solely as a decision criterion for the occurrence of insolvency. Nevertheless, our analysis shows that a supporting function is given and future research could further consider whether a separate consideration and classification of the new information identified in this way can provide further information on a change in the financial situation of a company. Furthermore, it has to be considered that our research is also limited in terms of the amount and scope of the data and should therefore be validated using a larger and balanced sample. Our methodology is suitable for this study to map informal changes and make them measurable, but it is subject to the limitation that no attention is paid to the order of content. Nevertheless, it is

also of interest in which order information is disclosed in a financial statement. Considering the sheer length of such a report, the question arises for future research to what extent the placement of information may be motivated by a company.

5. CONCLUSION

We were able to show that differences between solvent and insolvent companies can be made visible in the presentation of information in annual financial statements, which offer an approach to support the insolvency prognosis of companies. Nevertheless, it must be critically reflected that the financial situation and development alone are not necessarily decisive for a change in reporting. Nevertheless, our results concerning the selected data set also indicate that we should not expect extreme changes in consecutive financial statements. Concerning current research on the analysis of textual components of annual financial statements, we thus raise the question of whether an analysis of entire financial statements is instrumental in the assessment of the financial situation of companies or whether an analysis of the quantity difference is of more interest.

REFERENCES

Bankamp, S., and Muntermann, J. (2022). "Understanding the role of document representations in similarity measurement in finance and accounting". PACIS 2022 Proceedings.

Bankamp, S., Neuss, N., and Muntermann, J. (2021). "Crowd analysts vs. institutional analysts – A comparative study on content and opinion". Wirtschaftsinformatik 2021 Proceedings.

Bloomfield, R.J. (2002). "The "Incomplete Revelation Hypothesis" and financial reporting". Accounting Horizons 16, 3, 233-243.

Delivery Hero (2020). Annual report 2020. URL: https://ir.deliveryhero.com/download/companies/delivery/Annual%20Reports/DE000A2E4K43-JA-2020-EQ-E-01.pdf (visited on 02/22/2022).

Delivery Hero (2021). Annual report 2021. URL: https://ir.deliveryhero.com/download/companies/delivery/Annual%20Reports/DE000A2E4K43-JA-2021-EQ-E-01.pdf (visited on 02/12/2022).

Fromm, H., Wambsganss, T., and Söllner, M. (2019). "Towards a taxonomy of text mining features". Proceedings of the 27th European Conference on Information Systems, 1-12.

HGB (2021a). Handelsgesetzbuch §264 - Pflicht zur Aufstellung; Befreiung. URL: https://www.gesetze-im-internet.de/hgb/__264.html (visited on 11/15/2021).

HGB (2021b). Handelsgesetzbuch §267 - Umschreibung der Größenklassen. URL: https://www.gesetze-im-internet.de/hgb/__267.html (visited on 11/15/2021).

InsBekV (2021). Insolvenzbekanntmachungen. URL: https://www.insolvenzbekanntmachungen.de/ (visited on 12/10/2021).

Lang, M., and Stice-Lawrence, L. (2015). "Textual analysis and international financial reporting: Large sample evidence". Journal of Accounting and Economics, 60, 2-3, 110-135.

Le Maux, J., and Smaili, N. (2021). "Annual report readability and corporate bankruptcy". Journal of Applied Business Research, 37, 3, 73-80.

Lohmann, C., and Ohliger, T. (2020). "Bankruptcy prediction and the discriminatory power of annual reports: empirical evidence from financially distressed German companies". Journal of Business Economics, 90, 1, 137-172.

Myšková, R., and Hájek, P. (2020). "Mining risk-related sentiment in corporate annual reports and its effect on financial performance". Technological and Economic Development of Economy, 26, 6, 1422-1443.

scitkit-learn (2022). TfidfVectorizer. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html (visited on 03/20/2022).

Zhang, Y., Callan, J., and Minka, T. (2002). "Novelty and redundancy detection in adaptive filtering". In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. Ed. by K. Järvelin. New York, NY: ACM, p. 81.

EL RIESGO DE PERDER EL EMPLEO EN EMPRESAS DE RECIENTE CREACIÓN

Faustino Prieto
Universidad de Cantabria, Departamento de Economía, España
Faustino.prieto@unican.es

José María Sarabia
CUNEF Universidad, Departamento de Métodos Cuantitativos, España
josemaria.sarabia@cunef.edu

Vanesa Jordá
Universidad de Cantabria, Departamento de Economía, España
jordav@unican.es

RESUMEN

En este trabajo, se lleva a cabo un análisis exploratorio del riesgo de perder el empleo, de los trabajadores que forman parte de empresas (o establecimientos de las mismas) que se encuentran en sus primeros cinco años de vida. Para ello, utilizamos datos de tasa de destrucción de empleo de empresas en Estados Unidos, en el periodo 1978-2019, publicados por el United States Census Bureau; y evaluamos el efecto de la edad, período y cohorte de nacimiento sobre la evolución de dicha tasa de destrucción de empleo. Consideramos que los resultados del presente análisis son de utilidad en diversos campos, entre los que destacar su utilidad como guía en la toma de decisiones de trabajadores que opten por un nuevo empleo o de empresas de reciente creación que estén en proceso de contratación, así como de interés para entidades financieras y aseguradoras que comercialicen un seguro de protección de pagos.

1. INTRODUCCIÓN

¿Qué riesgo existe de que te despidan en una empresa de reciente creación, o en un establecimiento abierto hace poco tiempo dentro de una empresa ya establecida? Ésta es una cuestión de gran relevancia para trabajadores que desempeñan su profesión en dichas empresas o establecimientos, o para trabajadores en búsqueda de empleo en las mismas. También es de gran importancia para dichas empresas a la hora de planificar y organizar su personal contratado; así como para aquellas entidades financieras y aseguradoras que incluyen en su portafolio de productos y servicios un seguro de protección de pagos, normalmente vinculado a préstamos personales o hipotecarios contratados por dichos trabajadores.

Dicho despido puede producirse en empresas (o establecimientos de las mismas) que continúan su actividad, que deciden reajustar su plantilla y prescindir de parte de sus trabajadores, o en empresas que deciden no continuar con su actividad y cierran. Como referencia del segundo caso, destacar que de todos los establecimientos estadounidenses creados entre los años 1978 y 2014, sólo sobrevivieron a sus cinco primeros años de vida, aproximadamente entre el 41% y el 52% (Prieto et al, 2022).

Una variable que puede darnos información sobre dicho riesgo de despido es la tasa de destrucción de empleo anual de las empresas. Podemos encontrar trabajos de enorme interés que abordan el análisis de la evolución de dicha variable, por ejemplo: Den Haan et al. 2000, en el que se estudia la relación entre las fluctuaciones de la tasa de destrucción de empleo con la propagación de shock en la producción; Fuchs & Wey, 2010, en el que se examinan los determinantes de la creación y destrucción de empleo; Dosi et al. 2021, en el que se analiza la relación entre cambio tecnológico y empleo; entre otros. Muchos de ellos, centran su estudio en las empresas de reciente creación, por ejemplo: Kane, 2010; Haltiwanger, 2011; Davila et al. 2015; entre otros.

En este trabajo se analiza el efecto de la edad de los nuevos negocios, durante sus cinco primeros años de vida, en la evolución de dicha tasa de destrucción de empleo, como punto de referencia del riesgo de perder el empleo en los mismos. Dicho análisis se lleva a cabo con datos de tasas de destrucción de empleo de empresas estadounidenses, de edad entre 1 y 5 años, con año de nacimiento entre 1978 y 2014

(abordando, por tanto, los períodos comprendidos entre 1978 y 2019). Para ello se utiliza en primer lugar, un análisis gráfico de dichos datos, y a continuación, se considera un modelo de edad-periodo-cohorte ACP: Age-Period-Cohort (Yang & Land, 2013), en particular, la técnica denominada Median Polish (Tukey, 1977).

El contenido del presente trabajo es el siguiente. En la sección 2, se describe la base de datos considerada, así como la metodología utilizada para su análisis. En la sección 3, se muestran los resultados obtenidos. Finalmente, en la sección 4, se proporcionan las conclusiones del estudio realizado.

2. DATOS Y METODOLOGÍA

En este trabajo, se utilizó la base de datos "Business Dynamics Statistics" (BDS 2019), publicada por la Oficina del Censo estadounidense, en su versión de 30 de septiembre de 2021 (US Census Bureau, 2021). Dicha base de datos proporciona información de la tasa de destrucción de empleo de empresas y establecimientos estadounidenses en función de su edad, para los períodos 1978-2019. En concreto, para el presente análisis, se seleccionó la información correspondiente a empresas con 1 a 5 años de vida.

Como referencia, en la tabla 1, se muestran las tasas de destrucción de empleo (en tanto por ciento), correspondientes a los 7 primeros períodos (1978-1984) y a dichos cinco primeros años de edad (1-5 años), donde puede observarse que existen datos no disponibles en los cuatro primeros períodos (1978-1981), y donde se ha resaltado la primera cohorte disponible, nacida en 1978.

Tabla 1. Tasas de destrucción de empleo (%) de empresas estadounidenses, en sus primeros cinco años de vida, en los períodos 1978-1984. Fuente: US Census Bureau, BDS 2019.

	1978	1979	1980	1981	1982	1983	1984
1	28,554	27,218	31,967	28,008	29,627	30,901	30,667
2	NA	22,522	26,653	24,626	27,190	26,298	21,236
3	NA	NA	22,704	22,065	24,533	23,494	17,657
4	NA	NA	NA	19,283	22,548	22,824	17,262
5	NA	NA	NA	NA	20,957	22,306	16,572

Con respecto a la metodología seguida en el presente análisis exploratorio, en primer lugar, se realizó un análisis gráfico de la evolución de dichas tasas de desempleo a lo largo de los períodos y cohortes considerados, agrupando dichas tasas de desempleo en función de la edad de dichas empresas.

A continuación, se llevó a cabo un análisis de edad-periodo-cohorte (ACP: Age-Period-Cohort), siguiendo la metodología Median Polish (Tukey, 1977), que partiendo del siguiente modelo aditivo:

Tasa destrucción empleo = Común + Efecto Edad + Efecto Período + Residual

y en base a la siguiente expresión:

$$\gamma_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},\ i = 1, \ldots, 5;\ j = 1978, \ldots, 2019,$$

nos permitió estimar el efecto de la edad de la empresa (α_i) y el efecto del período considerado (β_j), en la evolución de dicha tasa de destrucción de empleo (γ_{ij}), así como de detectar valores extremos en los residuos obtenidos (ε_{ij}).

Para realizar dicha técnica Median Polish se utilizó el software R, en concreto, el comando: medpolish.

3. RESULTADOS

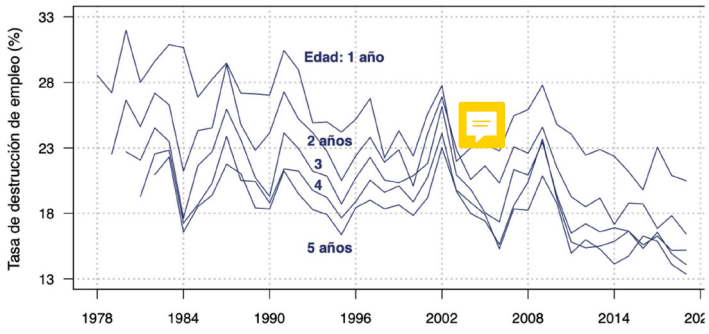
La figura 1 muestra la evolución de las tasas de destrucción de empleo (en tanto por ciento), de las empresas estadounidenses de nueva creación, agrupadas dichas empresas por su edad (de 1 a 5 años), en función del período considerado (arriba), y en función de la cohorte de nacimiento (abajo). Puede comprobarse visualmente, y en términos generales, que la evolución de dicha tasa parece seguir una tendencia decreciente para las cinco edades consideradas, tanto si consideramos el período como si consideramos la cohorte. También puede observarse un patrón decreciente en función de la edad, al aparecer en general cada línea por debajo de las correspondientes a edades inferiores a ella.

La figura 2 muestra la evolución de las tasas de destrucción de empleo (%) en función de la edad de las empresas durante sus primeros 5 años de vida, para los períodos 2000 y 2010 (izquierda) y las cohortes 2000 y 2010 (derecha). Puede comprobarse que dicha evolución fue muy diferente según el período o cohorte considerado, lo que confirma la necesidad del análisis edad-período-cohorte llevado a cabo.

La figura 3 muestra el efecto de la edad (izquierda) y el efecto del período (derecha), ambos en tanto por ciento, en la evolución de la tasa de destrucción de empleo (%), obtenidos utilizando la técnica Median Polish descrita anteriormente. Puede comprobarse que el efecto edad presenta un patrón decreciente, por lo que se confirma que, en base a la técnica utilizada, durante los cinco primeros años de vida, conforme una empresa de reciente creación va cumpliendo años, la tasa de destrucción de empleo sigue un patrón decreciente. También se confirma, en base al efecto período estimado, que dicha tasa de destrucción de empleo ha seguido una tendencia decreciente durante el intervalo de tiempo 1978-2019 considerado.

La figura 4 muestra los residuos obtenidos utilizando dicha técnica de Median Polish, indicando la tasa de destrucción de empleo que no ha podido ser explicada en base al efecto edad y al efecto período considerados. Para ello se han utilizado como colores extremos: el amarillo para los valores más bajos de dichos residuos, y el rojo para los valores más altos de los mismos.

Figura 1. Tasa de destrucción de empleo (%) de las empresas estadounidenses, durante sus primeros 5 años de vida, agrupadas dichas empresas por su edad (de 1 a 5 años), en función del período (arriba), y en función de la cohorte de nacimiento (abajo).



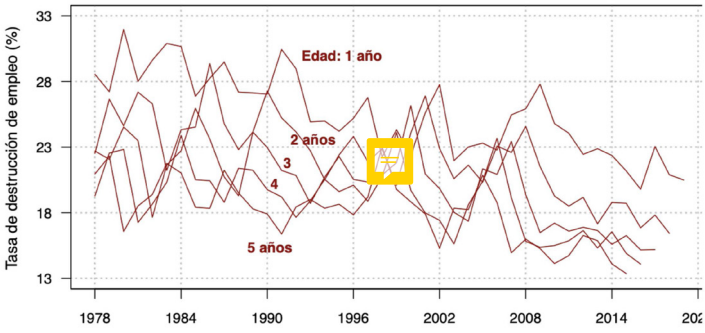


Figura 2. Tasa de destrucción de empleo (en tanto por ciento) de las empresas estadounidenses, durante sus primeros 5 años de vida, en función de la edad de dichas empresas (en años), para los períodos 2000 y 2010 (izquierda), y las cohortes 2000 y 2010 (derecha).

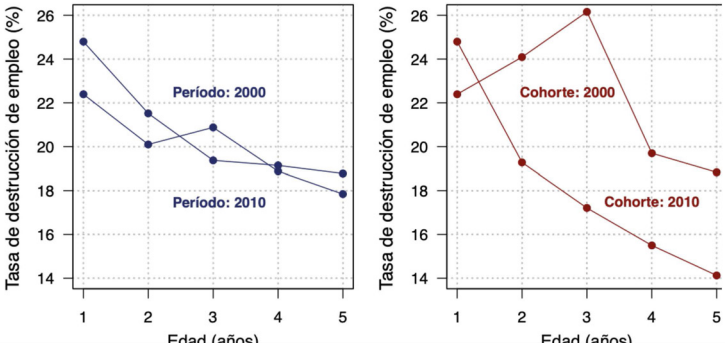


Figura 3. Efecto edad (izquierda), y efecto período (derecha), ambos en tanto por ciento, en la evolución de la tasa de destrucción de empleo, obtenidos utilizando la técnica Median Polish.

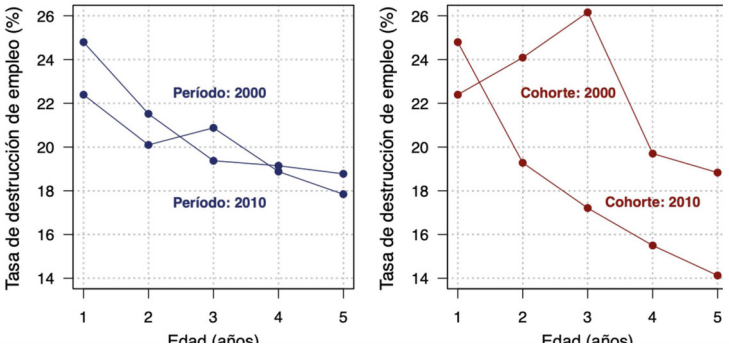
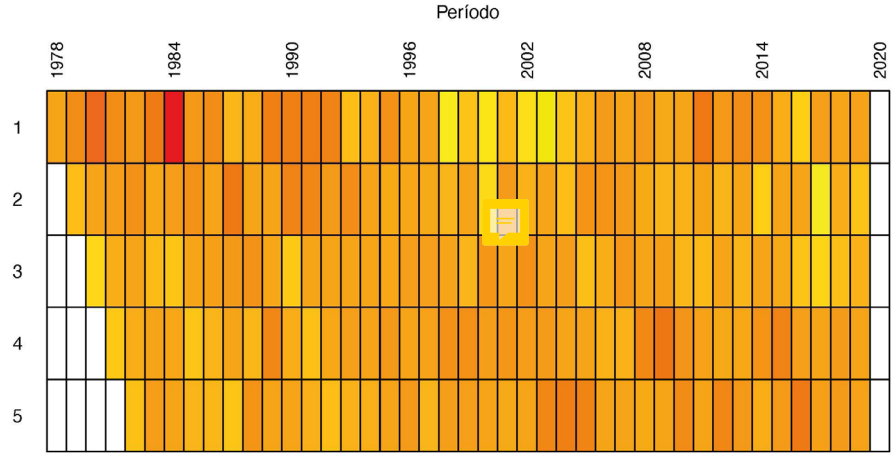


Figura 4. Residuos obtenidos utilizando la técnica Median Polish. Colores extremos: amarillo para valores más bajos, rojo para valores más altos.



4. CONCLUSIONES

Este trabajo presenta un análisis edad-periodo-cohorte (ACP), basado en la técnica de Median Polish, de la evolución de las tasas de destrucción de empleo de las empresas estadounidenses de reciente creación, durante sus primeros cinco años de vida, en los períodos comprendidos entre 1978 y 2019.

Hemos encontrado que, en dichas empresas, el efecto edad se traduce en una reducción de las tasas de destrucción de empleo conforme dichas empresas van cumpliendo años. Así mismo, hemos encontrado que el efecto período obtenido muestra que dichas tasas de destrucción de empleo han seguido una tendencia decreciente en el intervalo de tiempo 1978-2019 considerado.

AGRADECIMIENTOS

Esta publicación es parte del proyecto de I+D+i: PID2019-105986GB-C22, financiado por MCIN/AEI/10.13039/501100011033.

REFERENCIAS

Davila, A., Foster, G., He, X., and Shimizu, C. (2015). "The rise and fall of startups: Creation and destruction of revenue and jobs by young companies". Australian Journal of Management, 40,1, 6-35.

Den Haan, W.J., Ramey, G., and Watson, J. (2000). "Job destruction and propagation of shocks". American economic review, 90, 3, 482-498.

Dosi, G., Piva, M., Virgillito, M. E., and Vivarelli, M. (2021). "Embodied and disembodied technological change: The sectoral patterns of job-creation and job-destruction". Research Policy, 50, 4, 104199.

Fuchs, M., and Weyh, A. (2010). "The determinants of job creation and destruction: Plant-level evidence for Eastern and Western Germany". Empirica, 37, 4, 425-444.

Haltiwanger, J. (2011). "Job creation and firm dynamics in the US". Innovation Policy and the Economy, 12.

Kane, T.J. (2010). "The importance of startups in job creation and job destruction". Kauffman Foundation Research Series: Firm Formation and Economic Growth.

Prieto, F., Sarabia, J.M., and Calderín-Ojeda, E. (2022). "The risk of death in newborn businesses during the first years in market". Proceedings of the Royal Society A, 478, 20210952.

Tukey, J.W. (1977). Exploratory Data Analysis, Reading Massachusetts: Addison-Wesley.

US Census Bureau (2021). Business Dynamics Statistics Datasets, BDS 2019. https://www.census.gov/data/datasets/time-series/econ/bds/bds-datasets.html

Yang, Y., and Land, K.C. (2013). Age-period-cohort analysis: New models, methods, and empirical applications. Taylor & Francis.

UN ANÁLISIS DE SENSIBILIDAD BAYESIANA DESDE UN PUNTO DE VISTA MULTIVARIANTE CON APLICACIÓN EN PRINCIPIOS DE PRIMAS

Fabrizio Ruggeri
IMATI– CNR, Milano, Italy
fabrizio@mi.imati.cnr.it

Marta Sánchez-Sánchez
Universidad de Granada, Dpto. Estadística e Investigación Operativa, España
marta.sanchez@uca.es

Miguel Angel Sordo
Universidad de Cádiz, Dpto. Estadística e Investigación Operativa, España
mangel.sordo@uca.es

A. Suárez-Llorens
Universidad de Cádiz, Dpto. Estadística e Investigación Operativa, España
alfonso.suarez@uca.es

ABSTRACT

En este trabajo se aborda una metodología novedosa que permite introducir incertidumbre en el marco Bayesiano a través de definir clases de distribuciones a priori. Comprobamos que dichas clases nos conducen a la obtención de cotas superiores e inferiores para las primas Bayesianas o primas a posteriori. Cabe señalar, que esta nueva metodología se basa en las propiedades de los principios de prima para preservar el orden de las distribuciones a priori y posteriori a través de ordenaciones estocásticas multivariantes, donde usaremos densidades ponderadas para inducir incertidumbre sobre la información a priori. Finalmente, esta metodología tiene aplicaciones en análisis de sensibilidad Bayesiano y en los sistemas de tarificación Bonus-Malus.

1. INTRODUCCIÓN

Sea X una variable aleatoria representando un riesgo en actuariales. Un principio de prima asociado al riesgo X es una función, $H[X]$, que proyecta X en una cantidad no negativa determinista denominada prima. Esta prima es el precio que paga el asegurado para compensar a la aseguradora por asumir el riesgo subyacente. Existen numerosos principios de prima en la literatura, desde el simple valor medio del riesgo o prima neta a otros más complejos basados en la teoría de la utilidad esperada. Igualmente, existe un conjunto de propiedades deseables que determinan la coherencia de los distintos principios de prima tales como el principio de independencia, la carga de riesgo, la carga de riesgo no justificada, estafa, invarianza de escala, aditividad, subaditividad, superaditividad, monotonicidad, preservación de ordenaciones estocásticas, etc. Para más información véase Young (2004) y el Capítulo 2 en Denuit et al (2005).

Habitualmente, el riesgo X, con función de densidad, $f(x|\theta)$, depende de un parámetro n-dimensional, θ perteneciente a un espacio paramétrico, Θ; el cual caracteriza los riesgos individuales dentro de una misma clase o perfil de riesgo. Analizando el problema desde una aproximación Bayesiana, el valor del parámetro es cuantificable a partir de la información disponible; considerando dicho parámetro como una nueva variable aleatoria, π, conocida como distribución a priori del parámetro o función estructura. Si denotamos por $\pi(\theta)$ la densidad a priori y observando una muestra del riesgo aleatorio, $x = (x_1, ..., x_n)$, actualizamos la información a priori a través del Teorema de Bayes obteniendo la distribución a posteriori, π_x, cuya densidad viene dada por la expresión $\pi_x(\theta) = l(\theta|x)\pi(\theta) / m(x)$, donde $m(x)$ y $l(\theta|x)$ corresponden a la densidad marginal y a la función de verosimilitud de la muestra, respectivamente.

El planteamiento Bayesiano conduce a diferentes primas según el nivel de información. Dado el riesgo X y el principio de prima H, obtenemos la prima $H[X] = P_{R,H}(\theta)$, dependiente del parámetro θ denotada como prima individual o prima de Riesgo. Dada la dependencia del parámetro, desde un punto de vista Bayesiano $P_{R,H}(\theta)$ es también una variable aleatoria que simboliza un nuevo riesgo. Considerando un principio de prima H^* —no necesariamente igual a H— realizamos una nueva proyección y obtenemos $H^*[P_{R,H}(\pi)] = P_{C,H,H^*}(\pi)$, prima colectiva, la cual cuantifica

el riesgo inicial modelado por la distribución a priori. Análogamente, la prima Bayes, denotada por $H^*[P_{R,H}(\pi_x)] = P_{C,H,H^*}(\pi_x)$, cuantifica el riesgo una vez incorporada la experiencia muestral. Para más detalles véase Gómez-Déniz (1999) y Sánchez-Sánchez et al. (2019).

Por otra parte, la elección de la distribución a priori es una de las principales críticas y desventajas del modelo Bayesiano; esta elección es a veces subjetiva o simplemente busca cierta "tratabilidad" matemática, véase Gray y Pitts (2012). Una solución relevante en la literatura es sustituir la elección por una familia o clase de distribuciones a priori, denotada por Γ, que, igualmente, conducirá a una clase de distribuciones a posteriori, denotada por Γx, y, en el problema actuarial, a su vez conducirá a un conjunto de primas colectivas y Bayesianas. El estudio del rango de estas primas incide en la robustez del modelo y responde a las críticas en cuanto a la parcialidad y a la arbitrariedad de la elección de una única distribución a priori. Algunos ejemplos clásicos de clases de distribuciones a priori son las familias paramétricas, las clases contaminadas, las bandas de densidad, las bandas de distribución, etc.

El propósito principal de este trabajo es abordar el estudio de la robustez a partir de la clase de distribuciones a priori multivariante denominada banda ponderada, publicada en Ruggeri et al. (2019). Esta clase permite incorporar incertidumbre sobre la distribución a priori y, por ende, en los principios de primas. Finalmente, mostramos un ejemplo basado en un proceso de Poisson compuesto.

2. LA CLASE PONDERADA

A continuación, resumimos brevemente los conceptos teóricos en los cuales se basa la clase o banda ponderada definida en Ruggeri et al. (2019)

Definición 1. Sean X e Y dos vectores aleatorios n-dimensionales con funciones de densidad o masa de probabilidad (PDF) f y g, respectivamente.

a. X es menor que Y en el orden estocástico multivariante usual, denotado por $X \leq_{st} Y$, si $E[\phi(X)] \leq E[\phi(Y)]$, para toda función real creciente ϕ en $\mathbb{R}^n$ cuya esperanza existe.

b. X es menor que Y en el orden razón de verosimilitud multivariante, denotado por $X \leq_{lr} Y$, si $f(x)g(y) \leq f(x \wedge y)g(x \vee y)$, para todo x e y en $\mathbb{R}^n$.

Observación 1. La siguiente implicación es bien conocida en la literatura, ver Muller y Stoyan (2002) y Shaked y Shanthikumar (2007).

$$X \leq_{lr} Y \Rightarrow X \leq_{st} Y. \tag{1}$$

A continuación, recordamos el concepto de funciones MTP2.

Definición 2. Una función $\ell: \mathbb{R}^n \mapsto \mathbb{R}^+$ se dice que es multivariante totalmente positiva de orden 2 (MTP2), si ℓ satisface que $l(x)l(y) \leq l(x \wedge y)l(x \vee y)$.

Observación 2. Una función $\ell: \mathbb{R}^n \mapsto \mathbb{R}^+$, dos veces diferenciable, es MTP2 sí y solo si $\frac{\partial^2}{\partial x_i \partial x_j} \ln(\ell(\mathbf{x})) \geq \mathbf{0}$ **para todo** $\mathbf{i \neq j}$. Adicionalmente, un vector aleatorio n-dimensional es MTP2 si su densidad es MTP2 o, equivalentemente, si $X \leq_{lr} X$.

La construcción de la banda ponderada está basada en las funciones peso. Dada una prior multivariante π con función de densidad $\pi(\theta), \theta \in \Theta \subseteq \mathbb{R}^n$, se define la función peso, $w: \mathbb{R}^n \mapsto \mathbb{R}^+$ como una función no negativa tal que la esperanza $E^{\pi}[w(\theta)]$ es estrictamente positiva y finita. Entonces, un vector aleatorio ponderado, π_w, es un vector aleatorio que representa la incertidumbre en el conocimiento a priori a través de la perturbación asociada a w y cuya función de densidad viene dada por $\pi_w(\theta) = \pi(\theta)w(\theta)/E^{\pi}[w(\theta)]$.

Existen diversos criterios para elegir w a la hora de introducir la incertidumbre en el conocimiento a priori. En Ruggeri et al. (2019) consideran funciones ponderadas crecientes y decrecientes por dos razones. En primer lugar, porque representan una generalización de las distorsiones cóncavas y convexas usadas en las bandas distorsionadas en Arias et al. (2016). En segundo lugar, permiten comparar la distribución original con la ponderada, según vemos en el siguiente lema de Ruggeri et al. (2019).

Lema 1: Sea π una distribución a priori con densidad MTP2 y sea w una función peso creciente (decreciente). Entonces $\pi \leq_{lr} (\geq_{lr}) \pi_w$.

El Lema 1 es clave para definir una nueva clase de distribuciones a priori. Dadas w_1 y w_2 dos funciones ponderadas decreciente y creciente, respectivamente, y dada π una distribución a priori MTP2, se define la clase ponderada de π como:

$$\Gamma_{w1,w2,\pi} = \{\pi', \pi_{w1} \leq_{lr} \pi' \leq_{lr} \pi_{w2}\}. \quad (2)$$

Obviamente, como consecuencia del Lema 1, $\pi \in \Gamma_{w1,w2,\pi}$. Por tanto, la banda ponderada es una banda "de vecindad" de π donde las funciones a priori ponderadas relacionadas con w_1 y w_2 son las que alcanzan las cotas inferior y superior, respectivamente. Además, las distribuciones a posteriori heredan el ordenamiento en razón de verosimilitud bajo ciertas condiciones de regularidad. Por tanto, las distribuciones a posteriori de las cotas son igualmente cotas para las distribuciones a posteriori.

Proposición 1: Sea π una distribución específica a priori MTP2 y sea $\Gamma_{w1,w2,\pi}$ una banda ponderada asociada con π para w_1 y w_2. Dado los datos muestrales x, si la verosimilitud $l(\theta|x)$ es MTP2 en θ, entonces para todo $\pi' \in \Gamma_{w1,w2,\pi}$ obtenemos que $\pi_{w1x} \leq_{lr} \pi' \leq_{lr} \pi_{w2x}$, es decir, el orden se hereda a posteriori.

Para finalizar este epígrafe, destacamos que la banda ponderada cumple los requisitos recomendados en Berger (1994) para la definición de clases de distribuciones a priori. Por una parte, la elección y la interpretación son sencillas; por otra parte, usando diferentes métricas podemos cuantificar la incertidumbre a priori; y, además, el rango de la clase se calcula fácilmente a partir de las distribuciones de los extremos que definen la clase, como veremos en la aplicación. Para más detalles, ver Ruggeri et al. (2019).

3. APLICACIÓN: MODELO PROCESO DE POISSON COMPUESTO

Supongamos que el riesgo de un cliente sigue un clásico proceso de Poisson compuesto, $\sum_{i=1}^{N} X_i$, donde $\{X_i, i \geq 1\}$, es una secuencia de variables aleatorias i.i.d. con distribución común X, que a vez es independiente de *N*. La variable *N* representa el número de reclamaciones en un periodo, asumiremos con distribución Poisson, $N \sim P(\lambda)$. La variable *X* representa la cuantía de cada reclamación, asumiremos

con distribución exponencial, X ~ Exp(1/μ). Siguiendo el esquema Bayesiano, consideramos la distribución a priori del parámetro θ = (λ,μ), como $\pi = (\pi^1,\pi^2)$, un vector bivariante con marginales independientes distribuidas como exponencial y gamma inversa, $\pi^1(\lambda) \sim Exp(1/\alpha_1)$ y $\pi^2(\lambda) \sim Exp(2/\alpha_2)$, con hiperparámetros α_1 y α_2. Dada una muestra i.i.d de reclamaciones, $(n_1,\ldots,n_n)$, en n periodos asociada a otra muestra i.i.d de cuantías, $(x_1,\ldots,x_m)$, donde $\mathbf{m} = \sum_{i=1}^{n} \mathbf{n_i}$ es el número total de reclamaciones y teniendo en cuenta la independencia en el modelo de Poisson compuesto, se obtiene fácilmente la distribución a posteriori $\boldsymbol{\pi_x} = (\boldsymbol{\pi_x^1}, \boldsymbol{\pi_x^2})$ como un vector bivariante con marginales independientes distribuidas como gamma y gamma inversa, $\boldsymbol{\pi_x^1(\lambda) \sim Ga(\sum_{i=1}^{n} n_i + 1, n + 1/\alpha_2)}$ y $\boldsymbol{\pi_x^2(\mu) \sim IGa(2 + m, \alpha_2 + \sum_{i=1}^{m} x_i)}$.. Un planteamiento similar se puede ver en Boratynska (2021).

En este punto, inducimos incertidumbre en π a partir de las funciones peso $w_1(\lambda, \mu) = \lambda^{a1-1}\,\mu^{b1-1}$ y $w_2(\lambda, \mu) = \lambda^{a2-1}\,\mu^{b2-1}$, con $0 < a_1 < 1$, $0 < b_1 < 1$ y $0 < a_2 < 1$, $0 < b_2 < 2$, la cuales son decreciente y creciente, respectivamente; ambas mantienen la independencia de las marginales. Por tanto, se obtiene fácilmente que $\boldsymbol{\pi_{w_1} = (\pi_{w_1}^1, \pi_{w_1}^2)}$, $\boldsymbol{\pi_{w_2} = (\pi_{w_2}^1, \pi_{w_2}^2)}$, $\boldsymbol{\pi_{w_1,x} = (\pi_{w_1,x}^1, \pi_{w_1,x}^2)}$ y $\boldsymbol{\pi_{w_2,x} = (\pi_{w_2,x}^1, \pi_{w_2,x}^2)}$ son cuatro distribuciones bivariantes con marginales independientes tales que $\boldsymbol{\pi_{w_1}^1(\lambda) \sim Ga(a_1, 1/\alpha_1)}$, $\boldsymbol{\pi_{w_1}^2(\mu) \sim IGa(3 - b_1, \alpha_2)}$, $\boldsymbol{\pi_{w_2}^1(\lambda) \sim Ga(a_2, 1/\alpha_1)}$, $\boldsymbol{\pi_{w_2}^2(\mu) \sim IGa(3 - b_2, \alpha_2)}$, $\boldsymbol{\pi_{w_1,x}^1(\lambda) \sim Ga(\sum_{i=1}^{n} n_i + a_1, n + 1/\alpha_1)}$, $\boldsymbol{\pi_{w_1,x}^2(\mu) \sim IGa(3 + m - b_1, \alpha_2 + \sum_{i=1}^{m} x_i)}$, $\boldsymbol{\pi_{w_{2x}}^1(\lambda) \sim Ga(\sum_{i=1}^{n} n_i + a_2, n + 1/\alpha_1)}$ y $\boldsymbol{\pi_{w_2,x}^2(\mu) \sim IGa(3 + m - b_2, \alpha_2 + \sum_{i=1}^{m} x_i)}$.

Respecto a los principios de prima, a partir de ahora consideraremos H y H* como principios de prima neta. Por tanto, se obtiene que $H[X] = P_{R,H}(\theta) = \lambda\,\mu$ y, en consecuencia, las primas colectiva y Bayes se calculan como $H^*[P_{R,H}(\pi)] = E^{\pi}[\lambda\,\mu]$ y $H^*[P_{R,H}(\pi_x)] = E^{\pi}[\lambda\,\mu]$, respectivamente. Equivalentemente, las primas colectivas y Bayes se calculan para los extremos de la banda ponderada. La elección de la prima neta es por facilidad de cálculos, se obtienen resultados similares a partir de cualquier principio de prima no decreciente para θ = (λ, μ) .

Entonces, a partir de la definición de la clase distorsionada dada en (2), usando la propiedad de preservación del orden en razón de verosimilitud de la Proposición 1 y la implicación dada en (1), se obtiene fácilmente que todas las primas colectivas y Bayesianas quedarían acotadas entre las correspondientes asociadas a las

distribuciones ponderadas a priori y a posteriori, respectivamente. Por tanto, a partir de la preservación del valor esperado del orden estocástico multivariante, tenemos que para toda distribución a priori en la banda ponderada, $\pi' \in \Gamma_{w1,w2,\pi}$, se verifica que:

$$P_{C,H,H^*}(\pi_{w1}) \leq P_{C,H,H^*}(\pi') \leq P_{C,H,H^*}(\pi_{w2}) \text{ y } P_{B,H,H^*}(\pi_{w1,x}) \leq P_{B,H,H^*}(\pi') \leq P_{B,H,H^*}(\pi_{w2,x}).$$

Las desigualdades anteriores nos permiten estudiar la robustez del problema a partir de evaluar los valores extremos. A continuación, mostramos un ejemplo simulado para diferentes perfiles de clientes.

Supongamos una cartera en la que se produce de media una reclamación por cliente y por periodo, con cuantías medias por reclamación de 200 €. Basta pensar en una cartera de clientes de seguro de coches con cierta antigüedad, donde hay una alta siniestralidad y las reclamaciones no son excesivas debido a la baja tasación de los vehículos. Con el objetivo de hacer simulaciones fijamos los siguientes datos iniciales. En primer lugar, simulamos para el número de periodos n = 1, 3 y 5. Por otra parte, dado un cliente, fijamos constante el promedio de reclamaciones en cada periodo, *m*/*n*, e igualmente fijamos constante el promedio de las cuantías en cada periodo $\sum_{i=1}^{m} x_i/m$. La Tabla 1 muestra un resumen de los perfiles.

Tabla 1: Perfiles de clientes con promedios constantes en cada periodo.

n = 1, 3 y 5	Cliente 1 (C1)	Cliente 2 (C2)	Cliente 3 (C3)
m/*n*	1	1	2
$\sum_{i=1}^{m} x_i/m$	100	200	400

Consideramos los hiperparámetros $\alpha_1 = 1$ y $\alpha_2 = 200$. Inducimos incertidumbre en la banda con los valores $a_1 = 0.8$, $b_1 = 0.8$, $a_2 = 1.2$ y $b_2 = 1.2$. Un cálculo directo nos lleva a la Tabla 2.

Tabla 2: Primas netas colectiva y Bayesiana para la prior específica y para ambas cotas de la banda ponderada, para cada perfil de cliente y evaluando diferentes periodos.

	P_{C,H,H^*}	P_{B,H,H^*} (1)			P_{B,H,H^*} (2)			P_{B,H,H^*} (3)		
		n=1	n=3	n=5	n=1	n=3	n=5	n=1	n=3	n=5
π_{w1}	133.3	122.7	113.1	109.1	163.6	181.0	187.1	262.5	330.6	353.6
π	200.0	150.0	125.0	116.7	200.0	200.0	200.0	300.0	350.0	366.7
π_{w2}	300.0	183.3	138.2	124.7	244.4	221.1	213.8	342.9	370.6	380.2

Observamos en la Tabla 2 como la robustez mejora a medida que se mantienen los promedios de reclamaciones aumentando los periodos. Por el contrario, con pocos periodos la prima es sensible a la incertidumbre inducida.

4. APLICACIÓN: CÁLCULO DE PRIMAS BONUS-MALUS

En este epígrafe usamos la banda ponderada para estudiar la robustez del modelo de Poisson compuesto en un sistema Bonus-Malus. Para tal propósito, consideramos la siguiente familia de distribuciones a priori. Dado ε tal que $0 \leq \varepsilon \leq 1$, consideramos la distribución mixtura entre las dos cotas de la banda ponderada, $\pi_\varepsilon(\lambda,\mu) = (1-\varepsilon)\,\pi_{w1}(\lambda,\mu) + \varepsilon\pi_{w2}(\lambda,\mu)$.

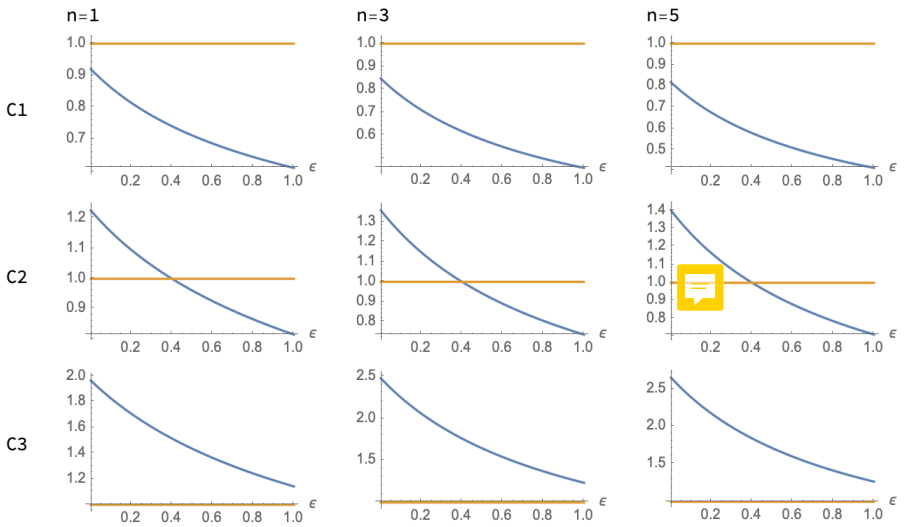
A través de la Proposición 2.10 de Ruggeri et al. (2019), sabemos que $\pi_\varepsilon \in \Gamma_{w1,w2,n}$. Para dicha distribución consideramos el cociente de laprima Bayesiana entre la prima colectiva, dado por $P_{B,H,H^*}(\pi_{\varepsilon,x})/P_{C,H,H^*}(\pi_{\varepsilon,})$, para más detalles ver Denuit et al. (2009) y Sarabia et al. (2004). Si dicho cociente es mayor o menor que la unidad, la experiencia provoca un grado de riesgo superior o inferior al colectivo. Si por el contrario fuese la unidad, la experiencia determina un grado de riesgo equivalente al colectivo; dicho grado de riesgo determina un cliente con una siniestralidad superior, inferior o igual al promedio colectivo. Tomando nuevamente *H* y *H** los principios de prima neta, un cálculo directo conduce a la siguiente expresión:

$$\frac{\mathbf{P_{B,H,H^*}(\pi_{\varepsilon,x})}}{\mathbf{P_{C,H,H^*}(\pi_{\varepsilon})}} = \frac{(\mathbf{1-\varepsilon})\,\mathbf{m_{w_1,x}}\,\mathbf{E^{\pi_{w_1,x}}[\lambda,\mu]} + \boldsymbol{\varepsilon}\,\mathbf{m_{w_2,x}}\,\mathbf{E^{\pi_{w_2,x}}[\lambda,\mu]}}{\Big((\mathbf{1-\varepsilon})\,\mathbf{E^{\pi_{w_1}}[\lambda,\mu]}+\boldsymbol{\varepsilon}\,\mathbf{E^{\pi_{w_1}}[\lambda,\mu]}\Big)\Big((\mathbf{1-\varepsilon})\mathbf{m_{w_1,x}}+\boldsymbol{\varepsilon}\mathbf{m_{w_2,x}}\Big)} \qquad (3)$$

donde $m_{w\_1,x}$ y $m_{w\_2,x}$ corresponden a las densidades marginales de las cotas de las distribuciones a posteriori $\pi_{w1,x}$ y $\pi_{w2,x}$, respectivamente. Observamos que el

cociente anterior varía para los distintos valores de ε. Dado ε = 0, obtendríamos la clasificación del riesgo para una distribución a priori con menos riesgo que la específica colectiva y dado ε = 1, para una con mayor riesgo. Por tanto, podríamos inferir que si para todos los valores de ε el cociente dado por (3) es siempre inferior a la unidad, la experiencia de dicho cliente lo clasifica de forma robusta como cliente poco conflictivo y susceptible a ser bonificado. Si, por el contrario, dicho cociente es siempre mayor que la unidad, el cliente es de forma robusta conflictivo y, por tanto, susceptible a ser penalizado. En cualquier otra situación, la experiencia del cliente podría clasificarlo como bueno o malo dependiendo de la incertidumbre introducida por la banda ponderada, por lo cual daría lugar a clientes poco robustos donde un mayor tamaño muestral o mayor experiencia sería necesaria antes de aplicarle cualquier beneficio o penalización.

Gráfica 1: Estudio de un sistema Bonus-Malus para diferentes perfiles de clientes. Por columnas veremos la evolución en el tiempo, es decir, considerando los distintos periodos de n. Por fila podemos encontrar los diferentes perfiles de los asegurados, en el mismo orden que en la tabla.



En la Gráfica 1 se muestra el cociente de primas dependiendo del valor de ε. Claramente los perfiles C1 y C3 son clientes robustos buenos y conflictivos, respectivamente, y los clientes C2 requieren una mayor observación.

4. CONCLUSIONES

En este trabajo hemos introducido una nueva metodología para inducir incertidumbre en la elección de una distribución a priori multivariante en el marco Bayesiano actuarial. Para tal propósito, hemos definido una clase de distribuciones a priori multivariante -banda ponderada- a partir de modificar la densidad a priori específica con funciones de peso. Hemos visto también que dicha clase es fácil de construir y nos permite analizar la sensibilidad del cálculo de las primas Bayesianas. Finalmente, basados en la banda ponderada, hemos propuesto un método gráfico y sencillo para evaluar los problemas de tarificación Bonus-Malus.

AGRADECIMIENTOS

Esta investigación ha sido financiada con el proyecto PID2020-116216GB-I00.

REFERENCIAS

Arias-Nicolás, J.P., Ruggeri, F., and Suárez-Llorens, A. (2016). "A gamma-minimax result in credibility theory". Bayesian Analysis, 11, 4, 1107-1136.

Berger, J. (1994). "An overview of robust Bayesian analysis (with discussion)". Test, 3, 5-124.

Boratynska, A. (2021). "Robust Bayesian insurance premium in a collective model with distorted priors under the generalized Bregman loss". Statistics in Transition. 22, 3, 123-140.

Denuit, M., Dhaene, J., Goovaerts, M.J., and Kaas, R. (2005). Actuarial Theory for Dependent Risks. John Wiley & Sons, New York.

Denuit, M., Marèchal, X., Pitrebois, S., and Walhin J.F. (2009). Actuarial Modelling of Claim Counts Risk Classification, Credibility and Bonus-Malus Systems. John Wiley & Sons, New York.

Gómez-Déniz, E., Hernández, A., and Vázquez-Polo, F.J. (1999). "The Esscher premium principle in risk theory: a Bayesian sensitivity study". Insurance: Mathematics and Economics", 25, 387-395.

Gray, R.J., and Pitts, S.M. (2012). Risk Modelling in General Insurance. Cambridge University Press.

Muller, M., and Stoyan, D. (2002). Comparison methods for stochastic models and risks. John Wiley & Sons, New York.

Ruggeri, F., Sánchez-Sánchez, M., Sordo, M.A., and Suárez-Llorens, A. (2019) "On a New Class of Multivariate Prior Distributions: Theory and Application in Reliability". Bayesian Analysis, 16, 1.

Sanchez-Sanchez, M., Sordo, M.A., Suarez-Lorens A., and Gómez-Déniz, E. (2019). "Deriving robust Bayesian premiums under bands of prior distributions with applications". Astin Bulletin, 49, 1, 147-168.

Sarabia, J., Gómez-Déniz, E., and Vázquez, F. (2004). "On the use of conditional specification models in claim count distributions: An application to bonus-malus systems". Astin Bulletin, 34, 85-89.

Shaked, M., and Shanthikumar, J.G. (2007). Stochastic orders. Series: Springer Series in Statistics, Springer.

Young, V.R. (2004). "Premium principles". In Encyclopedia of Actuarial Science. 1322-1331. John Wiley & Sons, New York.

CAPITAL FLOWS IN INTEGRATED CAPITAL MARKETS: THE MILA CASE

Juan David Vega Baquero
University of Barcelona, Dpt. of Econometrics, Statistics and Applied Economics,Spain
jvegabaq38@alumnes.ub.edu

Miguel Santolino
University of Barcelona, Dpt. of Econometrics, Statistics and Applied Economics,Spain
msantolino@ub.edu

ABSTRACT

The Feldstein and Horioka (1980) study on investment flows through the correlation of domestic saving and investment concluded that liberalization of capital markets does not necessarily lead to a movement of capital looking for a better allocation of resources, as classical theory would suggest. Ever since, literature has been prolific regarding this "puzzle", with arguments for and against this conclusion. This paper aims to analyze the issue from a different perspective. In recent years, the stock markets of Chile, Colombia, Mexico and Peru joined the Latin American Integrated Market (MILA), an agreement that allows investors in any of the participating markets to invest in the others as if they were investing locally. Traditional multivariate and compositional methods are used to assess the hypothesis of a potential flow of capital between markets generated by the creation of the joint market. As a result, it was not possible to find a change in the composition of the investment in the four markets produced by the creation of the joint market.

1. INTRODUCTION

The Feldstein and Horioka (1980) puzzle has caused great controversy since its publication. Many authors have argued in favor and against the results. The aim of this document is to use a more specific setting to assess the movement of capitals in fully integrated capital markets. More precisely, this approach will not enquire into the overall investment of a country, but on the portfolio investment and will analyze flows in the four markets forming the Latin American Integrated Market (MILA), an agreement that allows investors to trade stocks in the markets of the four participating countries locally.

Several methods have been proposed to assess the hypothesis in this case. Firstly, a time series approach through the estimation of multivariate models will be used to find a change in the dynamics of the composition of the MILA's market capitalization. Also, cross-sectional methods are used to assess the hypothesis from the compositional data perspective. As a result, multivariate time series models were not able to provide significant information to assess the hypothesis, while compositional methods in the cross-sectional side were able to provide consistent results, in favor of the Feldstein-Horioka puzzle.

2. THE FELDSTEIN-HORIOKA PUZZLE

The findings from Feldstein and Horioka (1980) marked the beginning of a discussion in economics: Is the liberalization of capital markets enough to generate net transfers of financial capital between countries? Classical theory would say yes, but the authors concluded that data do not agree. They used data on 21 OECD countries for the period 1960-1974 to assess the relationship between domestic saving and domestic investment. The idea behind is that with perfect mobility of capital, there should be no correlation between these two variables, since the investment decisions would respond to the opportunities in the global market. Indeed, they found that "international differences in domestic savings rates among major industrial countries have corresponded to almost equal differences in domestic investment rates" (Feldstein and Horioka, 1980, p. 328). As explained later by Ford and Horioka (2017), the liberalization of capital markets is not enough to

generate net transfers of financial capital between countries, so the integration of goods markets is also necessary to compensate the transfers of capital.

3. METHODOLOGY

Considering the nature of the dataset to be used, two approaches were followed. First, the series were modeled by means of vector autoregressive (VAR) models and second, compositional techniques for cross-sectional data were used.

3.1. Time series models and stationarity

Following Fuller (1996), a time series is a real function $x(t,\omega)$ for $t \in T$ (time) and $\omega \in \Omega$ (all possible realizations of the variable). Therefore, x is defined on $T \times \Omega$. For a fixed t, x is a random variable on a probability space. On the other hand, for a fixed ω, x is a function of time called a realization or sample function and is what can be observed in practice. Therefore, each one of the observations in the time series is a random variable itself and has its own distribution. When looking at the distribution of the observations over time (the sample function), it is a joint distribution function of all the individual random variables.

A time series is said to be weakly stationary if the expected value of x is constant for all t and the covariance matrix is only a function of the distance between the realizations and does not depend on t itself. In practice, the stationarity of a time series is tested using unit roots tests. The augmented Dickey-Fuller test (Said and Dickey, 1984) assesses the null hypothesis of unit roots in the characteristic equation of the time series, against the alternative of a stationary series. The idea behind is that all the roots of the characteristic equation of a stationary series should lie inside the unit circle.

Let consider the multivariate time series X_t of size m. For stationary multivariate time series, the vector autoregressive (VAR) model is defined as:

$$X_t = \sum_{i=1}^{p} \rho_i X_{t-i} + \epsilon_t \qquad (1)$$

where ρ_i is an $m \times n$ matrix of parameters and ϵ_t is the error term with multivariate normal distribution with means zero and with covariance matrix $\sum_{\epsilon}$, $N_m(0, \sum_{\epsilon})$. However, in real life many observed series are not stationary. Nevertheless, it is common that the changes in the variable follow a stationary process. In these cases, it is possible to estimate a VAR in differences to model the series. It is defined as:

$$\Delta X_t = \sum_{i=1}^{p} \rho_i \Delta X_{t-i} + \epsilon_t \tag{2}$$

Where Δ is the difference operator defined as $\Delta X_t = X_t - X_{t-1}$.

3.2. Compositional data

A composition is a multivariate series in which the variables carry only relative information (Pawlowsky- Glahn et al., 2011). It can be expressed as a vector $Y = (y_1, \cdots, y_n)$ where each y_i is a non-negative value and the $\sum_{j=1}^{n} y_j = \kappa$. The value of k is usually normalized to the unit ($\kappa = 1$). Formally, compositional data is defined in the sample space described by:

$$\mathcal{S}^n = \{Y \in R^n \big| y_j \geq 0, j = 1, \cdots, n, \sum\{_{j=1}^{n} y_j\} = 1 \tag{3}$$

The characteristics of S_n impose several restrictions when modeling the series. Therefore, Aitchison (1986) defined what is referred to as the Aitchison geometry. Thus, for two compositions $A, B \in S$ the perturbation operation $\oplus$ is defined as $A \oplus B = \left(a_1 \cdot b_1 / \sum_{j=1}^{n} a_j \cdot b_j, \cdots, a_n \cdot b_n / \sum_{j=1}^{n} a_j \cdot b_j\right)$ and the powering operation $\odot$ as $\odot A = \left(a_1^{\lambda} / \sum_{j=1}^{n} a_j^{\lambda}, \cdots, a_n^{\lambda} / \sum_{j=1}^{n} a_j^{\lambda}\right)$ for $\lambda \in \mathbb{R}$. The centered log ratio (clr) transformation is defined as clr(Y) = [ln(y_1/ g(Y)), ⋯ ,ln(y_n/$g(Y)$] with /$g(Y)$ $\left(\prod_{j=1}^{n} y_j\right)^{1/n}$. This transformation allows only for compositions with strictly positive values ($y_i > 0$), which imposes a limitation in the composition. Further discussion on the topic can be found in Aitchison (1986). On the other hand, even if the transformed elements can (in theory) take any value in $\mathbb{R}$, they still have the limitation of adding up to zero, because of the construction of the transformation. Therefore, the isometric log ratio (ilr) transformation is defined in the Aitchison geometry. In this case, an orthonormal basis $e = \{e_1, \ldots, e_{n-1}\}$ is used to create the matrix V with rows equal to $clr(e_j)$ and such that $V \cdot V' = I_{n-1}$ and $V' \cdot V = I_{n-1} - (1/n)1'_n, 1_n$, with I_n being the identity matrix of size n and 1_n an n-row vector of ones. Then, the irl transformation is defined as

$ilr(Y) = clr(Y) \cdot V$. For the specification of e, the binary partitions approach followed by Egozcue et al. (2003) can be used. With the transformed data, it is possible to use conventional statistical models and estimate a VAR. Therefore, the generic VAR model for a composition Y (after applying the ilr transformation) is defined as:

$$ilr(Y_t) = \sum_{i=1}^{p} \varrho_i ilr(Y_{t-i}) + \varepsilon_t \quad (4)$$

Like in the previous case, if the transformed series is not stationary, then it is possible to differentiate the series by applying the Δ operator and if the series in differences is stationary, then the model to estimate would be:

$$\Delta ilr(Y_t) = \sum_{i=1}^{p} \varrho_i \Delta ilr(Y_{t-i}) + \varepsilon_t \quad (5)$$

The VAR models will be assessed to see whether there is a change in the coefficients after the implementation of the agreement, meaning that there would be a potential movement of capitals within the markets.

3.3. Cross-sectional approach through compositional data

The compositional approach allows also for the use of other methodologies, analyzing the series as cross-sectional data instead of treating it as a time series. A first approach would be to estimate a Gaussian mixture model adjusted for compositional data as explained in Comas-Cufí et al. (2016). The idea behind this approach is that assuming the data comes from a mixture of three Gaussian distributions (one for each period analyzed), the model should assign the observations to these distributions following these breaks. If the model assigns the observations to the three distributions in a different way, then there is no evidence of a change in the composition of the MILA market in the periods analyzed and the Feldstein-Horioka (F-H) puzzle could not be rejected.

Additionally, it is possible to use different measurements of distance between compositions to check whether the entrance in force of the agreement produced a change in the composition of the joint market, meaning that there was a potential movement of capitals that would contradict the F-H puzzle. The first of these measurements is the Aitchison distance, defined by Aitchison (1986) as

$AD_{\Delta}(Y_t, \widehat{Y_t}) = \sqrt{(1/n)\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\ln(y_i/y_j) - \ln(\widehat{y_i}/\widehat{y_j})\right)^2}$. Also, following Thomas and Lovell (2014), it is possible to use measurements like the compositional Kullback-Leibler divergence, which measures how different two compositions are and is defined by Barceló-Vidal et al. (n.d.) as $KL_{\Delta}(Y_t, \widehat{Y_t}) = \sum_{i=1}^{n} \ln(y_i/\widehat{y_i})$. Thomas and Lovell (2014) also propose the cosine similarity, which ranges between –1 (completely opposite vectors) and 1 (completely proportional vectors), with 0 meaning completely orthogonal vectors. The cosine similarity is defined as $CS_{\Delta}(Y_t, \widehat{Y_t}) = (\sum_{i=1}^{n} y_i/\widehat{y_i})/(\sum_{i=1}^{n} y_i^2 \sum_{i=1}^{n} \widehat{y_i^2})$. For all three distance measurements, the distance will be taken each period with respect to the previous one, to see if there was an immediate shift in the composition of the MILA market that would indicate any potential movement of capitals at the moment of implementation of the agreement.

4. THE MILA MARKET

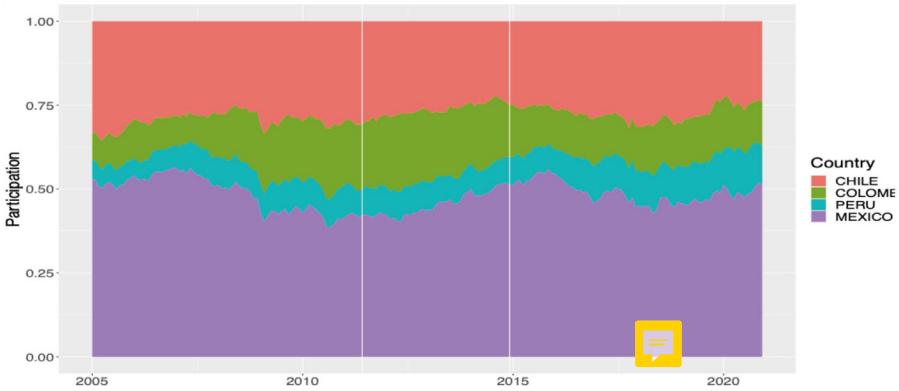
The Latin American Integrated Market (MILA) is an agreement signed by the stock exchanges from Chile, Colombia and Peru, which started operating as an interconnected market in 2011. In December 2014, Mexico joined the three founding members to form the current MILA. The four countries from the Pacific Alliance (a Latin American commerce bloc) signed the agreement which allows investors from any of the participating markets to invest in stocks from any of the exchanges. By the end of 2020, the market operated with more than 700 listed companies and had a capitalization above USD 770 billion. Operationally, the instruments are kept in the four separated exchanges, but they are interconnected for investors to be able to trade in any of the markets.

4.1. Dataset

The variable to be used is the monthly market capitalization of each stock market. The data is publicly available at the World Federation of Exchanges and is expressed in US dollars as common currency. The dataset is composed by the market capitalization of the four markets from January 2005 to December 2020. Figure 1 shows the composition formed by the four markets. There are two white lines dividing the graph: the first one corresponds to June 2011, when Chile, Colombia

and Peru joined the MILA and the second one to December 2014, when Mexico joined the three founding members and formed the current MILA. Therefore, the analysis will be concentrated in finding differences in the market capitalization of the four markets before and after the entrance in force of the agreement.

Figure 1. Composition of the market capitalization of MILA exchanges.



5. RESULTS

5.1. Times series modeling results

Once the dataset is defined, it is necessary to perform some diagnostic test in the series to determine the model to be estimated. The augmented Dickey-Fuller test assesses the null hypothesis of a unit root in the characteristic equation of the series, which implies that the series is non-stationary. None of the four series of market capitalization is stationary in levels. After differentiating the series one time, the test shows that the series are stationary, meaning that the model to be estimated is the one in equation 2. In the case of the compositions the situation is similar. The levels are not stationary, but the first differences are, so model in equation 4 is the one to estimate.

Later, it is necessary to determine how many lags are needed in each model. In this case, the information criteria to be used will be the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). Both use a likelihood method to define the fitness of a model to a sample. The AIC statistic is defined as $AIC = -2\ell + 2K$ where ℓ is the log-likelihood of the estimation and K the number of parameters in the model, whilst the BIC is defined as $BIC = -2\ell \ln(\ell) + K\ln(n)$ where n is the sample size. In both cases, a lower value of the statistic implies a better fit of the model.

Both the AIC and the BIC conclude that for the model with the differences of the market capitalization the more lags added the best. The number of lags was tested up to 36 and the AIC continued to decrease at each time. For the case of the compositional model the information criteria show that the model with only one lag has the best adjustment. For the sake of comparison, both models were estimated with one lag, considering that for the model with the differences of the market capitalization it was not possible to find an optimal number of lags and a parsimonious model is preferred. Therefore, the two models estimated are:

$$\Delta X_t = \rho \Delta X_{t-1} + \epsilon_t$$

$$\Delta ilr(Y_t) = \varrho \Delta \text{ilr}(Y_{t-1}) + \epsilon_t \qquad (6)$$

Most of the coefficients of the model for the market capitalization value in differences are non-significant. For the compositional model in differences the results are similar. Almost all the coefficients are non-significant. Altogether, these results show that it is not possible to find a relationship between the series and their lagged values that allows to analyze the hypothesis.

5.2. Cross-sectional analysis results

Figure 2 shows the results of the estimation of the Gaussian mixed model with three distributions. There seem to be two breaking points in the series, but they do not correspond to the entrance in force of the MILA market. Therefore, it is not possible to reject the F-H puzzle. Similar results are found with the measures of distance in figure 3. The left panel shows the Aitchison distance, which seems to

be close to zero along the whole period, without any important changes around the breaking points of interest. The middle panel shows the results for the compositional Kullback-Leibler divergence, which also remains near zero along the whole period. Finally, the left panel corresponds to the cosine similarity, which is close to one during the whole period, consistent with the previous results.

Figure 2. Gaussian mixed model results.

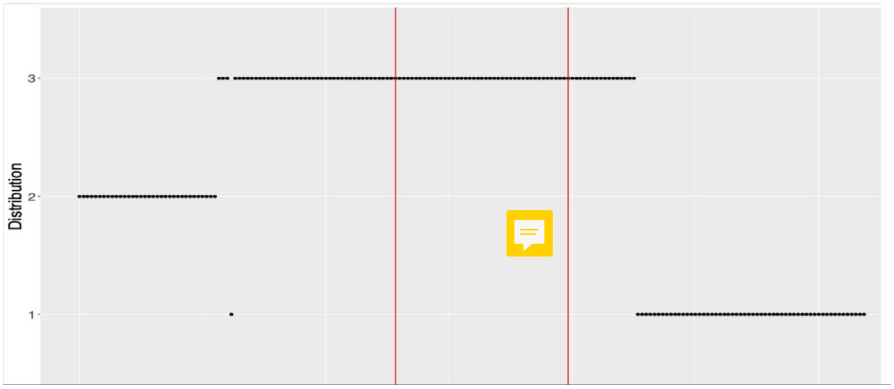
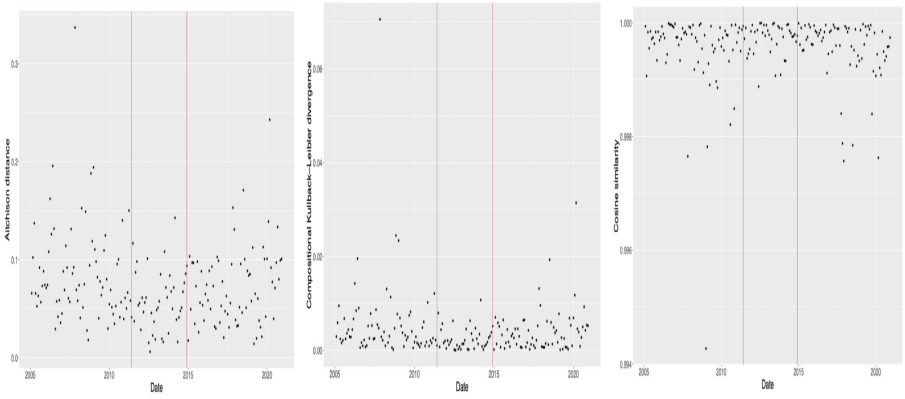


Figure 3. Aitchison distance results (left), Compositional Kullback-Leibler divergence results (center) and Cosine similarity results (right).



6. CONCLUSIONS

Different methods were applied to assess the Feldstein-Horioka hypothesis, according to which the liberalization of capital markets does not necessarily lead to a movement of capital between countries. The particular case of the Latin American Integrated Market (MILA) was used considering the specificity of the setting, in which only stock markets were liberalized allowing investors to buy shares in other markets. As a result, it was not possible to find any indication of capital flows within the markets after the implementation of the agreement, meaning that the Feldstein-Horioka puzzle holds also in this specific setting.

From the methodological point of view, multivariate time series models did not provide useful information to assess the hypothesis, as it was not possible to estimate a model with significant results. When analyzing the series as cross-sectional data, it was possible to use compositional methods to assess the hypothesis with consistent results. This sheds light on the potential of compositional methods for analyzing problems in which other approaches fail to provide significant results.

ACKNOWLEDGMENTS

The Spanish Ministry of Science and Innovation supported this study under grant PID2019- 105986GB-C21, as did the Secretaria d'Universitats i Recerca of the Catalan Government under grant 2020-PANDE-00074.

REFERENCES

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on statistics and applied probability. Chapman & Hall, London.

Barceló-Vidal, C., Bren, M., Martín-Fernández, J.A., and Pawlowsky-Glahn, V. (n.d.). "A measure of difference for compositional data based on measures of divergence". Unpublished manuscript. Available at: https://ima.udg.edu/~barcelo/index_archivos/ A_mesure_of_difference.pdf.

Comas-Cufí, M., Martín-Fernández, J.A., and Mateu-Figueras, G. (2016). "Log-ratio methods in mixture models for compositional data sets". SORT-Statistics and Operations Research Transactions, 1, 349-374.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). "Isometric logratio transformations for compositional data analysis". Mathematical Geology, 35, 3, 279-300.

Feldstein, M., and Horioka, C. (1980). "Domestic saving and international capital flows". Economic Journal, 90, 358, 314-329.

Ford, N., and Horioka, C. (2017). "The 'real' explanation of the Feldstein-Horioka puzzle". Applied Economics Letters, 24, 2, 95-97.

Fuller, W.A. (1996). Introduction to Statistical Time Series. John Wiley and Sons, New York, Second edition.

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2011). "Lecture notes on composi- tional data analysis". University of Girona.

Said, S.E., and Dickey, D.A. (1984). "Testing for unit roots in autoregressive-moving average models of unknown order". Biometrika, 71, 3, 599-607.

Thomas, P., and Lovell, D. (2014). "Compositional data analysis (CoDA) approaches to distance in information retrieval". Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '14. Association for Computing Machinery: Gold Coast, Queensland, Australia, 991-994

PARAMETRIC OUTSTANDING CLAIM PAYMENT COUNT MODELLING THROUGH A DYNAMIC CLAIM SCORE

Juan Sebastian Yanez
UQÀM, Département de Mathématiques, Canada
yanez.juan_sebastian@uqam.ca

Mathieu Pigeon
UQÀM, Département de Mathématiques, Canada
pigeon.mathieu2@uqam.ca

Jean-Philippe Boucher
UQÀM, Département de Mathématiques, Canada
boucher.jean_philippe@uqam.ca

ABSTRACT

By modelling reserves with micro-level models, individual claims information is better preserved and can be more easily handled in the fitting process. Some of the claim information is available immediately at the report date and remains known until the closure of the claim. However, other useful information changes unpredictably as claims develop, for example, the previously observed number of payments. In this paper, we seek to model payment counts in a discrete manner based on past information both in terms of claim characteristics and previous payment counts. We use a dynamic score that weighs the risk of the claim based on previous claim behaviour and that gets updated at the end of each discrete interval. In this paper's model we will also distinguish between the different types of payments. We evaluate our model by fitting it into a data set from a major Canadian insurance company. We will also discuss estimation procedures, make predictions, and compare the results with other models.

1. INTRODUCTION

In order to accurately predict the cost of future liabilities for open claims, practitioners and researchers have suggested several loss reserving models over the years. Over time, these models have changed a lot due to a significant increase in computing capacity, as well as in the quantity (and quality) of available data. While, in the past, models were always part of a collective framework, i.e. built for a data set aggregated by occurrence and development period, today we see a wide selection of models based on varying granularity of the underlying data set, ranging from raw data (micro-level) to aggregated data (macro-level).

Because of their granular structure, micro-level models can include more claim information in the modelling process than their aggregated counterparts. This information takes the form of covariates, which are of three types: static, time dynamic, and unpredictable time dynamic. Although time dynamic covariates change as time passes while static covariates remain fixed, both can be predicted with certainty at any point in time. In contrast, unpredictable time dynamic covariates are, as the name suggests, unpredictable. Thus, both static and time dynamic covariates can often be included in models in a more straightforward manner than unpredictable time dynamic covariates. Despite the uncertainty associated with the latter type of covariates, useful claim information can be extracted from them. Specifically, when modelling RBNS claims these covariates are abundant because a portion of the claim development has already been observed. Furthermore, few models that can handle this information have been implemented, namely Antonio et al. (2015), which considered including interchangeable states based on payment counts, and Pigeon et al. (2014) which made use of incurred losses. In this paper, we propose a new method that can handle an unpredictable time dynamic covariate in a discrete time interval framework.

For each of the open claims in the portfolio, we suggest using observed payments to improve the prediction of the future payments. Past payments are summarized using a score system that is updated at the end of a given discrete time interval with the new available information. Our discrete time scoring model can be implemented into any individual model that can predict payment counts at discrete intervals and that allows for the inclusion of covariates. This latter element is

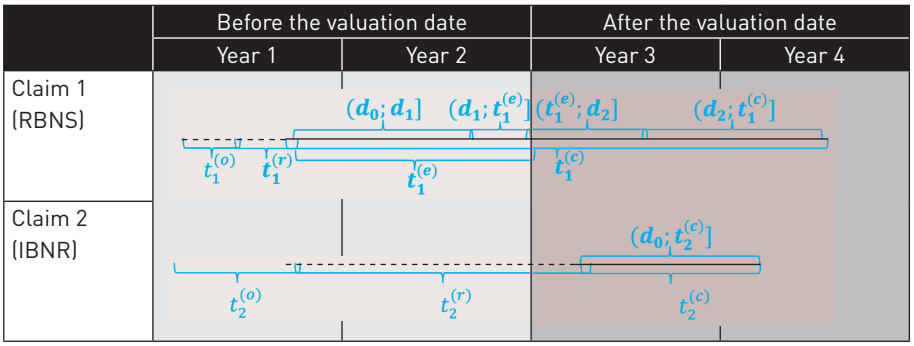
important because the claim score will be considered as a covariate. In particular, the frequency component in Yanez and Pigeon (2021) has both characteristics making it a candidate for the inclusion of this more intricate type of covariate.

The idea of calculating a score based on previous observations is not new to the actuarial literature. In fact, the model in this paper draws inspiration from the bonus-malus scoring system (BMS) developed for claim counts. Such a method was developed in Boucher and Inoussa (2014) where the authors summarized previous claim counts into a single numerical claim score. This model was further developed in Boucher (2022) a more compact and straightforward scoring system, called a Kappa-N model, was implemented. More recently Verschuren (2021) proposed a version of the model that incorporates the claim development of different product lines into the score system. In this work, we take inspiration from all these sources to introduce a similar dynamic claim scoring system into the micro-level reserving literature.

2. STATISTICAL FRAMEWORK

We show the typical development of a P&C claim in Figure 1. First, accident *i* occurs and we identify $t_i^{(o)}$, the occurrence delay, i.e., the delay between the beginning of the accident year and the exact accident date. There is an additional delay between the accident date and the reporting date denoted by $t_i^{(r)}$. After the accident has been reported, several payments may be made -- illustrated by dots in Figure 1, before the claim is closed after a final delay $t_i^{(c)}$. At the valuation date, claims can be split into two categories depending on their development. If the claim has not yet been reported we considered it Incurred But Not Reported, or IBNR, and if has been reported we consider it Reported But Not Settled, or RBNS. Furthermore, for RBNS claims we can compute $t_i^{(e)}$, the delay between the reporting date and the valuation date.

Figue 1. Development of two claims.



In a loss-reserving context, we first need to distinguish the status of each of the claims in the portfolio. Let $I = I^{(C)} \cup I^{(O)}$ be the set containing the claims available at the valuation date, where $I^{(C)}$ and $I^{(O)}$ are the subsets containing, respectively, the closed and the open (RBNS) claims. Let I^* be the set containing unreported claims (IBNR), which is unknown at the valuation date.

For each claim $i \in I$, the observation period, i.e., the period between the reporting date and the closure date (or the valuation date), is denoted by $(0;\tau_i]$, where $\tau_i = \min\{t_i^{(c)}, t_i^{(e)}\}$. Afterwards, the observation period, $(0;\tau_i]$, $i \in I$, can be divided into time intervals based on vector $d = [d_0, d_1, \ldots, d_K]$, where $d_k < d_{k+1}$, $d_0 = 0$ and $d_K > \max\{\tau_i\}$. For the sake of simplicity, we only consider an annual framework, i.e., d = [0,1,2, ...], but one could also consider a monthly or seasonal division. We suggest to base this decision on the expertise within the company, or on a cross-validation technique.

Furthermore, let $N_{i,k}$ be the number of payments for claim i, $i \in I$, taking place over the interval $(d_k, d_{k+1}]$. We can then identify the vector that holds these values at each interval as $N_i = [N_{i,0}, N_{i,1}, \ldots, N_{i,K-1}]$. To each $N_{i,K}$, we associate an exposure measure indicating how long claim i has been open over interval $(d_k, d_{k+1}]$. Thus, let $E_{i,k}$ be the exposure measure of claim i over the interval $(d_k, d_{k+1}]$:

$$E_{i,k} = \max\{\min\{\tau_i, d_{k+1}\} - d_k, 0\}$$

and $E_i = [E_{i,0}, E_{i,1}, \ldots, E_{i,K-1}]$..

At the reporting date, micro-level information from a claim becomes available in the form of a vector $X_i = [X_{i,1}, X_{i,2}, ..., X_{i,g}]$ of size g containing available static covariates, e.g., the region where the accident occurred, etc.

We can also identify a vector $Z_{i,k} = [Z_{i,k,1}, Z_{i,k,2}, ..., Z_{i,k,h}]$ of size h containing time dynamic covariates available at each interval $(d_k, d_{k+1}]$. In particular, this vector contains at least one covariate indicating the interval k with which $N_{i,K}$ is associated. Thus, this vector exists for reported claims, as well as for those that have not yet been reported (IBNR). For the latter, we define $Z^*_{i,K} = [d_k]$.

2.1 *A priori* distribution of the number of payments

For open claims, $i \in I^{[o]}$, we aim to predict the number of payments $N_{i,K}$, over the unobserved intervals after the valuation date $t_i^{(e)}$. We use the *a priori* information available at the reporting date (vectors X_i and $Z_{i,K}$), as well as the exposure $E_{i,K}$ before $t_i^{(e)}$. Commonly used approaches in a non-life-insurance context can be considered, such as generalized linear models (GLM). The expected value of $N_{i,K}$, conditionally to X_i, $N_{i,K}$ and $E_{i,K}$, is given by

$$\mu_{i,k} = E\left[N_{i,k} \middle| X_i, Z_{i,k}, E_{i,k}\right] = \left(E_{i,k}\right) g^{-1}\left(X_i \beta' + Z_{i,k} \theta'\right),$$

where $g^{-1}(.)$ is the inverse of the link function, β and $\theta$are, respectively, the parameter vectors of static and time dynamic covariates.

For claims that have occurred but have not been reported, $i \in I^*$, the report date occurs after valuation date, thus predictions must be made for all the intervals. Furthermore, instead of having access to the information contained in the vectors X_i and $Z_{i,K}$ we only have the information contained in $Z^*_{i,K}$.

2.2 a posteriori distribution of the number of payments

We suggested a method to model frequency payments at different intervals based on information from vectors X_i and $Z_{i,K}$, that respectively include static and time dynamic covariates. Now, we can focus on using unpredictable time dynamic

information through various measures. Let $\epsilon_{i,k}$ and $\eta_{i,k}$ be, respectively, the cumulative number of payments and exposure of claim i over the interval $(d_0, d_k]$:

$$\epsilon_{i,k} = \sum_{j=0}^{k-1} E_{i,j}, \qquad \eta_{i,k} = \sum_{j=0}^{k-1} N_{i,j}$$

We include the previously observed frequency in the mean parameter of claim i over interval $(d_k, d_{k+1}]$ in the following way:

$$\mu_{i,k} = E\left[N_{i,k} \middle| X_i, Z_{i,k}, \mathcal{H}_{i,k}\right] = (E_{i,k}) g^{-1}\left(X_i \boldsymbol{\beta}' + Z_{i,k} \boldsymbol{\theta}' + \gamma_1 \left(\eta_{i,k} / \epsilon_{i,k}\right)\right)$$

where $H_{i,K}$ is known development of claim i at time d_k, and γ_1 is the parameter associated with the new component. Then, we want to adjust the expected value of the frequency by incorporating a covariate that identifies payment-free periods in order to distinguish between claims that have been open for a longer or shorter period of time. Thus, as a claim develops, the frequency of payment-free periods may increase or reduce the expected value. This approach is inspired from the Kappa-N structure suggested by Boucher (2022). Let $K_{i,K}$ represent the total payment-free exposure observed over interval $(d_k, d_k]$ such that,

$$\kappa_{i,k} = \sum_{j=0}^{k-1} E_{i,j} \mathbb{1}(N_{i,j} = 0),$$

where $\mathbb{1}(.)$ is the indicator function. We can rewrite the mean parameter by incorporating both elements into a single claim:

$$\begin{aligned}
\mu_{i,k} = E\left[N_{i,k} \middle| X_i, Z_{i,k}, \mathcal{H}_{i,k}\right] &= (E_{i,k}) g^{-1}\left(X_i \boldsymbol{\beta}' + Z_{i,k} \boldsymbol{\theta}' + \gamma_0(-\kappa_{i,k}) + \gamma_1 \left(\eta_{i,k} / \epsilon_{i,k}\right)\right) \\
&= (E_{i,k}) g^{-1}\left(\mathbf{X}_i \boldsymbol{\beta}' + \mathbf{Z}_{i,k} \boldsymbol{\theta}' + \gamma_0 \left((-\kappa_{i,k}) + \left(\gamma_1 / \gamma_0\right)\left(\eta_{i,k} / \epsilon_{i,k}\right)\right)\right) \\
&= (E_{i,k}) g^{-1}\left(\mathbf{X}_i \boldsymbol{\beta}' + \mathbf{Z}_{i,k} \boldsymbol{\theta}' + \gamma_0 \left((-\kappa_{i,k}) + \psi \left(\eta_{i,k} / \epsilon_{i,k}\right)\right)\right) \\
&= (E_{i,k}) g^{-1}\left(\mathbf{X}_i \boldsymbol{\beta}' + \mathbf{Z}_{i,k} \boldsymbol{\theta}' + \gamma_0 \ell_{i,k}\right)
\end{aligned}$$

where k>0 and ψ is defined as the *jump-parameter*.

With this structure, we summarize past claim experience into a single claim score that will be updated at the end of each interval. Then, the mean parameter can identify claims that have higher chance of producing payments. Notice that $K_{i,K}$ is multiplied by –1 in order to better accommodate the negative impact that no-payment periods have on the claim score.

Note that the mean parameter is unbounded. This can be an issue because upper values of the mean parameter can become excessively large as we are including past frequency in our calculations and outliers are not uncommon. Thus, we suggest considering a maximum value for the claim score to avoid an overestimation of future payment counts. Moreover, the decreasing part of the measure, based on $K_{i,K}$, is bounded by the maximal duration of a claim, and is less prone to impacting excessively the prediction of the mean. Thus, the inclusion of a minimal value for the mean parameter is less suitable. Finally, when we look into new claims, no past history has been previously observed, and we cannot include the dynamic claim score measure. Thus, by setting the initial value of the claim score to 0, all predictions of the mean parameter are based only on the other covariates available at the report date. We suggest obtaining a claim score such that:

$$\ell_{i,k} = \begin{cases} \min\left\{\left(-\kappa_{i,k} + \psi\left(\eta_{i,k}/\epsilon_{i,k}\right)\right).\ell_{max}\right\}, & \text{for k} = 1,2,\ldots \\ 0, & \text{for k} = 0 \end{cases}$$

One should note that the claim score for claim *i* is updated at the end of each interval k based on information up to the end of the previous interval k – 1. As such, it is possible to identify which claims are more likely to produce payments derived from past information summarized by the value of the claim score at any given time.

3. NUMERICAL RESULTS

For our numerical analysis, we considered a data set from a Canadian insurance company. The data set contains information from 57,593 claims about Accident Benefits (AB) coverage, i.e., no-fault benefits for accidents where the insured, or a third party, was injured or killed in a car accident. Micro-level information is incorporated in the modelling process in the form of categorical static covariates. In

order to evaluate the performance of our model, we chose to set the valuation date in order to split the data set into a training and an evaluation set. Diving more deeply into the number of payments from the data set, which is the focus of this paper, we grouped payments into three categories: 1) Medical, 2) Disability and 3) Expenses.

Our first objective was to assess the performance of the inclusion of the claim score $\ell_{i,k}$ into frequency models, in terms of goodness-of-fit. First, we compared the likelihood, the AIC and BIC of two versions of our models. The first version included $\ell_{i,k}$ as a covariate and the second version did not. We present these results in Table 1 for the RBNS models. As shown in this table, the inclusion of the claim scores provides better results in terms of BIC and AIC across all models and all types of payments.

Table 1. AIC and BIC of RBNS models with and without the claim score

		AIC		BIC	
Model	Payment Type	with	without	with	without
NB	Medical	232,270	236,794	233,085	237,149
	Disability	81,099	84,576	81,464	84,931
	Expenses	122,865	123,964	123,230	124,320
POI	Medical	331,810	358,342	332,165	358,688
	Disability	203,585	240,730	203,940	241,077
	Expenses	160,370	164,481	160,725	164,828

Then, we compared the best performing model (the one that uses the Negative Binomial distribution) to other models in the literature. However, because most models directly predict the total cost of payments rather than payment counts, we decided to compare the total cost instead. Thus, we added a severity model to our dynamic score frequency model. We tested popular distributions such as the Gamma, log-Normal and inverse normal distributions. We found that fitting distributions for each type of payment separately and including the claim score as a covariate were satisfactory, and the Gamma distribution was chosen for this numerical analysis. As for the comparative distributions, we chose two collective generalized linear models, based on the quasi-Poisson distribution and the Gamma distribution. We also considered the individual model by Yanez and Pigeon

(2021), which served as a comparative baseline for the inclusion of dynamic claim scores. Table 2 contains the results of 10,000 simulations of each described model.

We notice that all the models yield satisfactory results in terms of the 95% and the 99% VaRs as the values are higher than the observed value. The two collective models (Gamma and over-dispersed Poisson) have a mean that is lower than the observed value, but their standard deviation is higher than the individual models.

Table 2. Results of the total reserve predictions

	Mean	Std. Dev	95% VaR	99% VaR
GLM Gamma	143,604,545	7,969,902	156,696,768	162,534,340
GLM ODP	145,171,862	6,565,836	156,112,224	161,073,565
3comp RBNS	145,459,940	3,636,952	151,546,231	154,130,897
3comp Total	149,620,225	3,678,054	155,830,382	158,291,786
Score RBNS	137,509,168	4,785,344	145,451,829	148,969,071
Score Total	142,852,107	4,842,791	150,950,931	154,342,708
Obs. RBNS	141,830,856			
Obs. Total	147,308,364			

Furthermore, because the 95% and the 99% Values-at-Risk of individual models are lower than the collective models but higher than the observed value, the latter approaches are preferable. As for the comparison between both individual approaches, we notice that the mean of the total reserve is lower for the dynamic score model however through a higher standard deviation, the 95% and the 99% VaRs become lower than the model that does not make use of the claim score. This further increases the accuracy of the model by providing values over the observed reserve. Again, this shows an overall numerical preference for the model in this paper over the one suggested in Yanez and Pigeon (2021).

ACKNOWLEDGMENTS

This research was financially supported by The Co-operators Research Chair in Actuarial Risk Analysis (CARA) .

REFERENCES

Antonio, K., Godecharle, E., and Van Oirbeek, R. (2016). "A multi-state approach and flexible payment distributions for micro-level reserving in general insurance". Available at SSRN 2777467.

Boucher, J.P., and Inoussa, R. (2014). "A posteriori ratemaking with panel data". ASTIN Bulletin: The Journal of the IAA, 44, 3, 587-612.

Boucher, J.P. (2022). "Bonus-Malus Scale Models: Creating Artificial Past Claims History". To be published in Annals of Actuarial Science.

Pigeon, M., Antonio, K., and Denuit, M. (2014). "Individual loss reserving using paid–incurred data". Insurance: Mathematics and Economics, 58, 121-131.

Verschuren, R.M. (2021). "Predictive claim scores for dynamic multi-product risk classification in insurance". ASTIN Bulletin: The Journal of the IAA, 51, 1, 1-25.

Yanez, J.S., and Pigeon, M. (2021). "Micro-level parametric duration-frequency-severity modeling for outstanding claim payments". Insurance: Mathematics and Economics, 98, 106-119.

MODELING NEGATIVE RATES

Miklós Arató
Eötvös Loránd University, Department of Probability Theory and Statistics, Hungary
miklos.arato@ttk.elte.hu

Dalma Tóth-Lakits
Eötvös Loránd University, Department of Probability Theory and Statistic,.Hungary
dalma.tothlakits@gmail.com

ABSTRACT

In our research the focus is on the modeling and calibration of a new phenomenon, the negative forward rates. After the theoretical background is presented in the beginning of the article, we show how forward rates should be modeled using continuous random fields. Furthermore, we demonstrate how the maximum likelihood estimations of the parameters can be derived. We offer an efficient way to simulate the Kennedy field for modeling the forward rate. The small amount of data which is enough to establish probability 1 estimations may be surprising.

1. INTRODUCTION

The appearance of the negative rates is a new phenomenon in the financial world, which raises quite a few problems in the field of mathematical modeling. In the current environment, the modeling of forward rates with random fields is proved to be a relevant and interesting issue as well, which can represent a completely new approach. The goal of our research is to derive analytical estimations of the parameters and to use machine learning algorithms as well, hence we can compare the achieved results.

The emergence of negative interest rates has highlighted the possibility of modeling financial processes with continuous Gaussian fields. The popularity of fractional processes has not decreased in the recent years and the tools of artificial intelligence are also becoming more popular and widespread.

Therefore, the primary goal of this article is to determine the maximum likelihood estimations and the estimations with probability 1 for the Kennedy fields introduced later. Additionally, we also paid close attention to the simulation of the fields the fastest way possible. The aim of our research is to be able to compare these estimations obtained with machine learning tools with the estimations obtained with classical analytical tools, although in this article, we basically cover the theory of the estimations.

2. MAXIMUM LIKELIHOOD ESTIMATION

In this section the theoretical background of the maximum likelihood estimations are presented.

Definition (Gaussian process). *Let* (Ω, A, P) *be a probability space, T is a parameter set. Then* $\xi\colon\ \Omega \times \mathrm{T} \to \mathbb{R}$ *is a Gaussian process, if for any* $n \in \mathbb{N}$ *and* $c_1, \ldots c_n \in \mathbb{R}$, $t_1, \ldots t_n \in T$, $\sum_{i=1}^{n} c_i \xi_{t_i}$ *is normally distributed. Then P is called a Gaussian measure in* (Ω, F_ξ). *For simplicity we can assume that* $A = F_\xi$. *The expected value and the standard deviation of the Gaussian process are marked as follows*

$$m(t) = E_\xi(t),\ B(s,t) = cov(\xi(s),\xi(t)).$$

It is well known that two Gaussian measures are either equivalent or orthogonal.

2.1 The case of different expected values

Let $\xi\colon\ \Omega \times \mathrm{T} \to \mathbb{R}$ be a Gaussian process. Let the expected value of the Gaussian process under the measure P be 0 and the expected value under the measure P_1 be m.

$$E_P\, \xi(t) = 0,\ E_{P1}\, \xi(t) = m(t),\ t \in T$$

Let *U* denote the linear space of the variables of the following shape

$$\sum_{i=1}^{n} c_i \xi_{t_i}, \qquad n \in \mathbb{N}, c_1, \dots, c_n \in \mathbb{R}, t_1, \dots, t_n \in T$$

Also take the following scalar product

$$< u, v \geq \int_{\Omega} uv dP$$

Finally, $\overline{U}$ denotes the Hilbert space obtained by closing *U.*

Rozanov showed that for different expected values, the Radon-Nikodym derivative can be calculated as follows. *[Rozanov (1971)]*

Theorem (Rozanov). *The P and P_1 measures are equivalent if and only if there exists an* $\eta \in \overline{U}$ *with which the expected value of the process can be written in the form below*

$$m(t) = \int_{\Omega} \xi(t)\eta(t) dP, \qquad t \in T$$

In the case of equivalence, the Radon-Nikodym derivative of the two measures is

$$\frac{dP_1}{dP} = e^{\eta - \frac{<\eta,\eta>}{2}}$$

A simple consequence of this theorem is the following statement by Arató. *[Arató(1997)]*

Theorem (Arató). *Let* ξ: $\Omega \times T \to \mathbb{R}$ *be a Gaussian process. Let the expected value of the Gaussian process under the measure P be 0 and the expected value under the measure Pm be m · a(t). The P and Pm measures are equivalent if and only if there exists an* $\eta \in \overline{U}$ *for which*

$$a(t) = \int_{\Omega} \xi(t)\eta(t) dP, \qquad t \in T$$

In the case of equivalence, the Radon-Nikodym derivative of the two measures is

$$\frac{dP_m}{dP} = e^{m\eta - \frac{m^2 < \eta,\eta >}{2}}.$$

Theorem (Maximum likelihood estimation) *Let ξ(t) be a Gaussian process. Then using the notations of the previous statement, the maximum likelihood estimation of* m is

$$\widehat{m} = \frac{\eta}{< \eta, \eta >}$$

This estimation is normally distributed and unbiased and the standard deviation is

$$D^2_{P_m} \widehat{m} = \frac{1}{< \eta, \eta >}$$

Proof. The shape of the estimation can be derived immediately from the Radon-Nikodym derivative. To determine the expected value and the standard deviation, calculate the next expected value when X ~ N(0, σ^2).

$$E(X^k e^{mX}) = \int_{-\infty}^{\infty} x^k e^{mx} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{x^2}{2\sigma^2}} dx = e^{\frac{m^2\sigma^2}{2}} \int_{-\infty}^{\infty} x^k \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-m\sigma^2)^2}{2\sigma^2}} dx$$

By calculating the first two moments we get the following values: if k = 1, then $E\,(Xe^{mX}) = m\sigma^2\, e^{m2\sigma 2/2}$ and if k=2 then $E\,(X^2\, e^{mX}) = (\sigma^2 + m^2\sigma^4)\, e^{m2\sigma 2/2}$. For the expected value we obtain the following

$$E_{P_m} \widehat{m} = \frac{1}{< \eta, \eta >} \int_\Omega \eta dP_m = \frac{1}{< \eta, \eta >} \int_\Omega \eta e^{m\eta - \frac{m^2 < \eta,\eta >}{2}} dP =$$

$$= \frac{1}{< \eta, \eta >} e^{-\frac{m^2 < \eta,\eta >}{2}} m < \eta, \eta > \; e^{\frac{m^2 < \eta,\eta >}{2}} = m$$

Similarly to the first moment, we can derive the second moment.

$$E_{P_m} \widehat{m}^2 = \frac{1}{< \eta, \eta >^2} \int_\Omega \eta^2 dP_m = \frac{1}{< \eta, \eta >^2} \int_\Omega \eta^2 e^{m\eta - \frac{m^2 < \eta,\eta >}{2}} dP =$$

$$= \frac{1}{< \eta, \eta >^2} e^{\frac{m^2 < \eta,\eta >}{2}} (< \eta, \eta > + m^2 < \eta, \eta >^2) e^{\frac{m^2 < \eta,\eta >}{2}} = \frac{1}{< \eta, \eta >} + m^2$$

From these derivations the standard deviation can be deduced immediately. Since these are Gaussian processes, the normality is obvious.

2.2 THE CASE OF CONSTANT EXPECTED VALUE

Theorem. *In the case where the expected value of the Gaussian process is constant, then a(t) = 1 for every t ∈ T. Let F = σ{ξ_t – ξ_s, s, t ∈ T). Let us fix a t_0 ∈ T point and h(ξ) = E_P ξ(t_0) | F) Let us assume that D_2 {ξ(t_0) – h(ξ)} > 0 Then the maximum likelihood estimation of m is*

$$\widetilde{m} = \xi(t_0) - h(\xi).$$

Proof. The proof is based on the idea of G. Shevchenko using the law of total expectation and $E_P\left(\widetilde{m}\big(\xi(t_0) - h(\xi)\big)\right) = D_P^2\,\widetilde{m}$. .

$$E_P\big(\widetilde{m}\xi(t)\big) = E_P\left(\widetilde{m}\big(\xi(t) - \xi(t_0) + \xi(t_0) - h(\xi) + h(\xi)\big)\right) =$$

$$= E_P\left(\widetilde{m}\left(\xi(t) - \xi(t_0) + h(\xi)\right)\right) + D_P^2\widetilde{m} =$$

$$= E_P(E_P\left(\left(\widetilde{m}\big(\xi(t) - \xi(t_0) + h(\xi)\big) \mid \mathcal{F}\right)\right) + D_P^2\widetilde{m} = D_P^2\widetilde{m}$$

Based on the previous derivations the maximum likelihood estimation of the expected value is the following

$$\widehat{m} = \frac{\widetilde{m}/D_P^2\widetilde{m}}{D_P^2(\widetilde{m}/D_P^2\widetilde{m})}.$$

2.3 Some simple examples

The well-known results of various stochastic processes which are often used to model financial processes immediately follow from the previous statements.

For example, let us observe a Gaussian process with *m* expected value and the same covariance as the Wiener process on the interval [*a*, *b*], where a > 0 and *a* < *b*. In this case, the maximum likelihood estimation of the Wiener process is the value of the process at the starting point: $\widehat{m} = \xi(a)$.

On the other hand, a stationary Ornstein-Uhlenbeck process in the [0, *T*] interval can also be observed. We know the value of $\lambda > 0$ in advance and the expected value and the covariance of the process

$$E_{P_m}\xi(t) = m$$

$$cov_{P_m}\big(\xi(s),\xi(t)\big) = \sigma^2 e^{-\lambda|t-s|}, \qquad s,t \in [0,T]$$

Therefore, the following covariances can be easily determined

$$E_P\big(\xi(0)\xi(t)\big) = \sigma^2 e^{-\lambda t}, E_P\big(\xi(t)\xi(T)\big) = \sigma^2 e^{-\lambda(T-t)}$$

$$E_P\Big(\int_0^T \xi(s)ds\xi(t)\Big) = \sigma^2\frac{2 - e^{-\lambda t} - e^{-\lambda(T-t)}}{\lambda}.$$

Taking advantage of the fact that in this case the maximum likelihood estimation is unbiased, we get the well-known Grenander formula:

$$\widehat{m} = \frac{\xi(0) + \xi(T) + \lambda\int_0^T \xi(s)ds}{2 + \lambda T}.$$

Norres et all investigated the case of the fractional Brownian motion with unknown trend. Consequently, the $\xi(t)$ process is observed in the [0, T_1] interval, where $\xi(t) - \alpha t$ is a fractional Brownian motion with Hurst parameter *H* under the measure *P*α *[Norros (1999)]* We also got the same result with the notation $\eta = \int_0^T w(T,s)d\xi(s)$, hence the Radon-Nikodym derivative can be calculated as follows

$$\frac{dP_\alpha}{dP} = e^{\alpha\eta - \frac{\alpha^2 D_P^2\eta}{2}}.$$

Furthermore, the unbiased maximum likelihood estimation of α is

$$\hat{\alpha} = \frac{\eta}{D_P^2\eta}.$$

3. FORWARD RATES

Let {$F(s, t)$, $0 \le s \le t$} denote the forward rates while $P(s,t) = e^{-\int_s^t F(s,u)du}$ the price of the bond paying a unit in time t at time s.

A classic scheme for modeling the forward rates is the framework created by Heath, Jarrow and Morton. *[Heath (1992)]* However, Kennedy looked at the forward rates as continuous Gaussian random fields with special expected value and covariance matrix. This model is consistent with the initial term structure, and it incorporates stochastic spot rates naturally. A sufficient condition on the drift surface is derived to ensure that the discounted bond prices of zero-coupon bonds are martingales. The parameters of the field described as follows: $\sigma, \lambda \ge 0$, $\mu \ge \lambda\backslash 2$, ν. *[Kennedy (1994)] [Kennedy (1997)]*

The expected value of the Gaussian random field is

$$\mu(s,t) = \nu - \sigma^2\left(\frac{1}{\mu} - e^{-\mu(t-s)}\left(\frac{1}{\mu} - \frac{1}{\lambda-\mu}\right) + e^{-\lambda(t-s)}\frac{1}{\lambda-\mu}\right)$$

and the covariance is

$$cov(F(s_1,t_1), F(s_2,t_2) = \Gamma(s_1,t_1,s_2,t_2) = \sigma^2 e^{\lambda \min(s_1,s_2)+(2\mu-\lambda)\min(t_1,t_2)-\mu(t_1+t_2)}$$

We can notice that the $F(s, s + t)$ field is an Ornstein-Uhlenbeck process in the variable s therefore

$$cov\big(F(s_1, s_1+t), F(s_2, s_2+t)\big) = \sigma^2 e^{-\lambda t} e^{-\mu|s_1-s_2|}$$

This means that if we can observe the $F(s, s + t)$ process on an interval according to s for some value t, then $\sigma^2 e^{-\lambda t}\mu$ is determined with probability 1. If we can do this for 2 different t value, then $\sigma^2\mu$ and λ are defined with probability 1.

If we look at another covariance from the field

$$cov\left(F\left(\frac{\log s_1}{\lambda}, t\right), F\left(\frac{logs_2}{\lambda}, t\right)\right) = \sigma^2 e^{-\lambda t}\min(s_1, s_2)$$

Which means that $\sigma^2 e^{-\lambda t}$is defined with probability 1, therefore also σ^2 and μ are defined with probability 1.

3.1 Simulation of the kennedy field

Simulating a multidimensional normally distributed vector is extremely slow due to the size of the covariance matrix. A much more effective, simpler, and faster way is if we notice that if $W(x, y)$ is a Brownian sheet, then $\mu(s,t) + \sigma e^{-\mu t} W(e^{\lambda s}, e^{(2\mu-\lambda)t})$ is a Kennedy field with the appropriate covariance structure.

The question is how can we generate a Brownian sheet at the (x_1, y_1) points the fastest way possible, where the division is $0 = x_0 < x_1 < \cdots < x_n$, $0 = y_0 < y_1 < \cdots < y_m$.

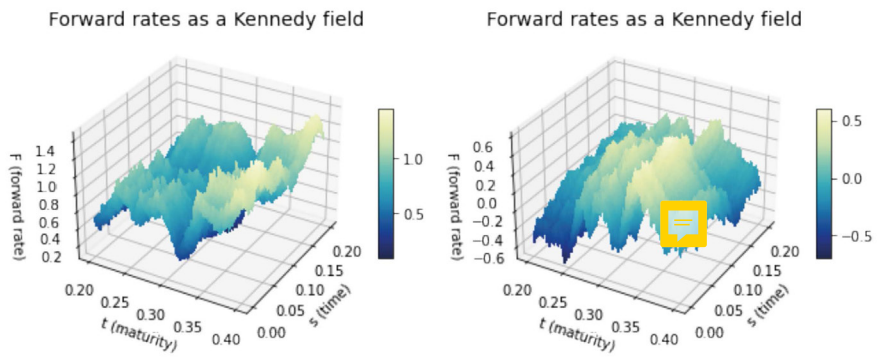
Let us take independent random variables with $N(0, (x_i - x_{i-1})(y_j - y_{j-1}))$ distributions and denote them with $\xi(i,j)$. Accordingly, the Brownian sheet can be written in the following form

$$W(x_i, y_j) = \sum_{k=1}^{i} \sum_{l=1}^{j} \xi(k, l)$$

Hence, the upcoming matrix operation should be coded the fastest way possible to achieve the des

$$A \to B: B(i,j) = \sum_{k=1}^{i} \sum_{l=1}^{j} A(k, l).$$

Figure 1. Simulated Kennedy fields



We also assume of the previous {F (s, t), (s, t) ∈ T} Kennedy field that

$$\{(a,t): b_1 \le t \le b_2\} \subseteq T \subseteq \{(s,t): s \le t, a \le s, b_1 \le t \le b_2\}.$$

Let ξ (s, t) mark the following random field

$$\xi(s,t) = F(s,t) + \sigma^2 \left(\frac{1}{\mu} - e^{-\mu(t-s)}\left(\frac{1}{\mu} - \frac{1}{\lambda-\mu}\right) + e^{-\lambda(t-s)}\frac{1}{\lambda-\mu}\right).$$

In this case, the maximum likelihood estimation of parameter ν is

$$\hat{\nu} = \frac{\frac{e^{\lambda b_1}}{\mu}\xi(a,b_1) + \frac{e^{\lambda b_2}}{\mu-\lambda}\xi(a,b_2) + \int_{b_1}^{b_2} e^{\lambda v}\xi(a,v)dv}{e^{\lambda b_2}\left(\frac{1}{\lambda} + \frac{1}{\mu-\lambda}\right) + e^{\lambda b_1}\left(\frac{1}{\mu} - \frac{1}{\lambda}\right)}.$$

4. CONCLUSION

Overall, the conclusion is that a new result has been achieved in the estimation of the parameters of the Kennedy field. This can be a great starting step for the investigation of other, more complicated models. The further goal of our research is to calibrate the parameters of the negative interest rate models with machine learning algorithms; therefore, we can compare them with the analytical estimations derived in this article.

ACKNOWLEDGMENTS

This work is supported by the KDP-2021 program and the ELTE TKP 2021-NKTA-62 funding scheme of the Ministry of Innovation and Technology from the source of the National Research, Development and Innovation Fund.

REFERENCES

Arató, N.M. (1997). "Mean estimation of Brownian sheet". Computers Mathematics with Applications, 33, 8, 13-25.

Heath, D.C., Jarrow, R.A., and Morton A. (1992). "Bond pricing and term structure of interest rates: A new methodology for contingent claims valuation". Econometrica, 60, 77-105.

Kennedy, D.P. (1994). "The term structure of interest rates as a Gaussian random fields". Mathematical Finance, 4, 247-258.

Kennedy, D.P. (1997). "The characterizing Gaussian models of the term structure of interest rates". Mathematical Finance, 7, 107-118.

Norros, I., Valkeila, E., and Virtamo, J. (1999). "An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions". Bernoulli, 5, 4, 571-587.

Rozanov, Ju.A. (1971). "infinite-dimensional Gaussian distributions: Proceedings (Proceedings of the Steklov Institute of Mathematics number 108 (1968))". American Mathematical Society, Providence, Rhode Island.

LISTADO DE AUTORES

Agarwal, Ruchi	Management Development Institute (MDI), India
Alcañiz, Manuela	Universitat de Barcelona, RiskCenter-IREA, Spain
Anaya Luque, David	Universitat de Barcelona, PhD Student, Spain
Arató, Miklós	Eötvös Loránd University, Hungary
Atance, David	Universidad de Alcalá, Spain
Ayuso, Mercedes	Universitat de Barcelona, Riskcenter-IREA, Spain
Barranco Chamorro, Inmaculada	Universidad de Sevilla, Spain
Belles Sampera, Jaume	Riskcenter-IREA, Grupo Catalana Occidente, Spain
Bello, Alfonso José	Universidad de Cádiz, Spain
Bermúdez, Lluís	Universitat de Barcelona, RiskCenter-IREA, Spain
Bolancé, Catalina	Universitat de Barcelona, RiskCenter-IREA, España
Boucher, Jean-Phillippe	UQÀM, Canada
Calderín Ojeda, Enrique	University of Melbourne, Australia
Carracedo, Patricia	Universitat Politècnica de València, Spain
Carrillo-García, Rosa M.	Universidad de Sevilla, Spain
Castaño Martínez, Antonia	Universidad de Cádiz, Spain
Céspedes, Luis E.	Zurich Insurance, Spain
Coia, Vicenzo	University of British Columbia, Vancouver, Canada
Debón, Ana	Universitat Politècnica de València, Spain
Devesa Carpio, Enrique	Universidad de Valencia, Spain
Devesa Carpio, Mar	Universidad de Valencia, Spain
Domínguez Fabián, Inmaculada	Universidad de Extremadura, Spain
Durban, Marc	Universitat Politècnica de Catalunya, Spain
Encinas Goenechea, Borja	Universidad de Extremadura, Spain
Estévez, Marc	Universitat de Barcelona, RiskCenter-IREA, Spain

Fernández Fontelo, Amanda	Universitat Autònoma de Barcelona, Spain
Gabarró, Joaquim	Universitat Politècnica de Catalunya, Spain
Gómez Déniz, Emilio	ULPGC, España
Guillen, Montserrat	Universitat de Barcelona, RiskCenter-IREA, Spain
Jordá, Vanesa	Universidad de Cantabria, Spain
Karlis, Dimitris	University of Economics and Business, Greece
Krysiak, Urszula	University of Applied Sciences in Ciechanow, Poland
Krysiak, Zbigniew	Warsaw School of Economics, Poland
Malczewski, Adrian	University of Social Sciences, Poland
Malczewski, Krzysztof	University of Life Sciences, Poland
Martos Ramírez, Albert	Universitat de Barcelona, RiskCenter-IREA, España
Meneu Gaya, Robert	Universidad de Valencia, Spain
Moriña, David	Universitat de Barcelona, RiskCenter-IREA, Spain
Mulero, Julio	Universidad de Alicante, Spain
Nießner, Tobias	University of Goettingen, Germany
Pigeon, Mathieu	UQÀM, Canada
Pigueiras, Gema	Universidad de Cádiz, Spain
Prieto, Faustino	Universidad de Cantabria, Spain
Puig, Pedro	Universitat Autònoma de Barcelona, Spain
Ramos, Carmen D.	Universidad de Cádiz, Spain
Ruggeri, Fabrizio	IMATI– CNR, Italy
Sánchez Sánchez, Marta	Universidad de Granada, Spain
Santolino, Miguel	Universitat de Barcelona, RiskCenter-IREA, Spain
Sarabia, Jose María	CUNEF Universidad, Spain
Schumann, Matthias	University of Goettingen, Germany
Sordo, Miguel Ángel	Universidad de Cádiz, Spain
Suárez Llorens, Alfonso	Universidad de Cádiz, Spain
Tóth-Lakits, Dalma	Eötvös Loránd University, Hungary
Vega Baquero, Juan David	Universitat de Barcelona, RiskCenter-IREA, Spain
Vidal Llana, Xenxo	Universitat de Barcelona, RiskCenter-IREA, Spain
Yanez, Juan Sebastian	UQÀM, Canada

You can download the digital version in the
Documentation Center
www.fundacionmapfre.org/documentacion

Fundación **MAPFRE**

Fundación **MAPFRE**

23/C

Paseo de Recoletos, 23
28004 Madrid (España)
www.fundacionmapfre.org

P.V.P.: 20 €
ISBN 978-84-9844-758-3