PROGRAMME and BOOK of ABSTRACTS

JOCLAD 2022

Porto, 18 July 2022

Associação Portuguesa de Classificação e Análise de Dados

U.PORTO
FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

Programme and Book of Abstracts

XXIX Meeting of the Portuguese Association for Classification and Data Analysis (CLAD)

18 July 2022

Porto, Portugal

https://joclad2022.fep.up.pt/

Sponsors

Associação Portuguesa de Classificação e Análise de Dados
Faculdade de Economia da Universidade do Porto
Banco de Portugal
Instituto Nacional de Estatística

Programme and Book of Abstracts

XXIX Meeting of the Portuguese Association for Classification and Data Analysis (JOCLAD 2022)

Editors: Paula Brito, Carlos Ferreira, Conceição Rocha, Pedro Campos, Sónia Dias

Publisher: CLAD

Preface

Welcome to JOCLAD 2022, the XXIX Meeting of the Portuguese Association for Classification and Data Analysis. JOCLAD aims at bringing together CLAD members, and all researchers and practitioners interested in Data Science. This year, JOCLAD is organised in a reduced format, on just one day, as Portugal is hosting the 17th conference of the International Federation of Classification Societies, IFCS 2022 – *Classification and Data Science in the Digital Age*. The (mini)-JOCLAD 2022 take place at the Faculty of Economics of the University of Porto, on the afternoon of 18 July, preceding IFCS 2022.

A keynote lecture is addressed by Berthold Lausen (University of Essex, UK), whom we warmly thank for joining us at JOCLAD 2022. The programme also includes Special Sessions, organized by the Bank of Portugal and Statistics Portugal, CLAD institutional sponsors, to whom we are much indebted for their lasting and generous support. JOCLAD 2022 is also the opportunity to honour CLAD's new emeritus members, Prof. Helena Bacelar Nicolau and Prof. Paulo Gomes.

We thank the chairs of the sessions, for their active collaboration. Finally, we thank the Faculty of Economics of the University of Porto for granting the conditions for JOCLAD 2022, and all people there whose work contributed to making it come true.

Last but not least, a big "thank you" goes to all participants, for your support to JOCLAD 2022, and for helping us to make this meeting a success. We hope to meet you again for the JOCLAD 2023!

Porto, July 2022

Paula Brito
CLAD President

Organization

CLAD President

Paula Brito (Fac. Economia, Univ. Porto & INESC TEC)

Local Organizing Committee

Carlos Ferreira (DEGEIT, Univ. Aveiro & IEETA)
Conceição Rocha (INESC TEC)
Pedro Campos (Fac. Economia, Univ. Porto & INESC TEC & Statistics Portugal)
Sónia Dias (ESTG, IPVC & INESC TEC)

Contents

Programme Overview

Monday, 18 July

13:30	Registration	FEP Hall
14:00	**Opening**	Room 118
14:15	**Keynote Lecture**	Room 118
15:15	**Session to Honour New CLAD Emeritus Members**	Room 118
16:15	Coffee Break	FEP Hall
16:45	**Bank of Portugal Session**	Room 118
17:45	**Statistics Portugal Session**	Room 118
18:25	**Closing**	Room 118

Programme

Monday, 18 July

13:30 Registration - FEP Hall

14:00 **Opening Session** Room 118

14:15 **Keynote Lecture** Room 118

Bootstrap evaluation of unsupervised statistical learning and applications
Berthold Lausen, p. 5

Chair: José G. Dias

15:15 **Session to Honour New CLAD Emeritus Members** Room 118

Helena Bacelar Nicolau, p. 9

Paulo Gomes, p. 11

Chair: Paula Brito

16:15 **Coffee Break**

16:45 **Thematic Session - Bank of Portugal** Room 118
Chair: Fernanda Sousa

16:45 **Complex made simple: Visualizing the business groups' database using Tom Sawyer**
Diogo Silva, Tiago Pinho Pereira, p. 15

17:05 **Live data – A game changer for official statistics**
Elena Bucea, Carlos Figueira, p. 17

17:25 **Give us data... and we give you back a wordless story**
Carolina Rocha, Mariana Oliveira, p. 19

Monday, 18 July

17:45 **Thematic Session - Statistics Portugal** Room 118
Chair: Pedro Campos

17:45 **Classification of CPP – Application of a multilayer neural network**
Almiro Moreira, Ana Carmona, M. Conceição Ferreira, David Santos, Rui Alves, p. 23

18:05 **Estimates on completed buildings for the Indicators System of Urban Operations – Exploring with ML methodologies**
André Sousa, António Portugal, Inês Sá, Pedro Cunha, Sara Cerdeira, p. 25

18:25 **Closing Session** Room 118

Abstracts

Keynote Lecture

18 July, 14:15 - 15:15

Bootstrap evaluation of unsupervised statistical learning and applications

Berthold Lausen

University of Essex

The talk provides an overview on estimating confidence limits of estimated clusters by unsupervised learning methods. The problem is illustrated by the task to estimate hierarchical clustering by distance data of high dimensional observations without class labels (unsupervised). Ultrametric and additive tree metric are mathematical models of phylogenetic inference or hierarchical clustering of distance data. Paul O. Degens introduced an additive measurement error model for distance data which was used by [2, 4] to develop a three objects variance estimator which provides a point estimate of the variance parameter without estimating the overall phylogenetic tree as an ultrametric or additive tree. Estimating the unknown location parameter, ultrametric or dendrogram, the three-objects variance estimator is used to compute parametric bootstrap estimates of the probability to observe the estimated clusters [3]. The approach is applied in the context of user segmentation based on online behavioural data [1]. We discuss and compare the parametric bootstrap approach with other recent suggestions of confidence limits for point estimates by unsupervised statistical learning.

References

[1] S. Hadjiantoni, H. Yang, Y. Long, R. Petraityte, and B. Lausen. User segmentation based on online behavioural data via ensemble predictions and clustering. In P. Brito, J.G. Dias, B. Lausen, A. Montanari, and R. Nugent, editors, *Classification and Data Science in the Digital Age, Proc. IFCS 2022*. Springer, 2023.

[2] B. Lausen and P.O. Degens. Variance estimation and the reconstruction of phylogenies. In *Die Klassifikation und ihr Umfeld*. Indeks Verlag Frankfurt am Main, 1986.

[3] B. Lausen and P.O. Degens. Bootstrap evaluation in hierarchical cluster analysis. In E. Diday, editor, *Data Analysis and Informatics V*. North-Holland, Amsterdam, 1988.

[4] B. Lausen and P.O. Degens. Evaluation of the reconstruction of phylogenies with dna-dna-hybridization data. In H.-H. Bock, editor, *Classification and Related Methods of Data Analysis: Proc. First Conference of the International Federation of Classification Societies (IFCS)*, pages 367–374. North-Holland, Amsterdam, 1988.

Special Session
New CLAD Emeritus Members

18 July, 15:15 - 15:45

Helena Bacelar Nicolau

Professor Helena Bacelar Nicolau holds a PhD in Sciences – Statistics and Probability (1960-1965) – from the University of Lisbon, having obtained later (1969-1972) a 3ème Cycle degree in Mathematical Statistics and Data Analysis from the University of Paris VI (France).
Her academic career was centred at the University of Lisbon, in a network perspective with collaborations in several Faculties of this University, such as the Faculty of Sciences (until 1989), the Faculty of Pharmacy (1989-2005), the Faculty of Medicine (2004-2011), where she was Director of its Biomathematics Laboratory, Faculty of Psychology and Educational Sciences, where she co-founded (with Professor Fernando Nicolau) and coordinated (1988-2011) the Laboratory of Statistics and Data Analysis (LEAD) and retired as Full Professor in 2011. She has also taught at other institutions, such as the University of Aveiro, creating local knowledge in Multivariate Data Analysis, with emphasis on Classification methods. Her non-academic professional activity has also been related to Statistics and Data Science, namely through the consulting activity in the company Datascience Consultores of which she is partner since 2000.
She is a founding member and was the first President of the Association for Classification and Data Analysis, from 1994 to 2000, establishing CLAD's association with the International Federation of Classification Societies (IFCS) in 1995. In this International Federation she has held several positions, such as: Council Member of the IFCS Finance Committee in the periods 1995-2000 and 2010-2014; Member of the IFCS Election Council (2005-2009) and Chair of the IFCS Education Committee (2002-2005). She founded the CLAD Bulletin, first edited in 1996 and has been co-editor, since the first volume in 2013, of the publication Classification and Data Analysis – Methods and Applications (CLADMAp). She integrated, in a regular and continuous way, the Scientific Committee of the annual Jornadas de Classificação e Análise de Dados (JOCLAD), having presented the paper "Modeling in Hierarchical Classificatory Analysis" at the first one, held at the Faculty of Psychology and Education Sciences of the University of Lisbon, in 1993. She is Honorary President of Fernando Nicolau Award.

18 July, 15:45 - 16:15

Paulo Gomes

Professor Paulo Gomes obtained a Master's Degree in Statistics and Operational Research (1983-1984) at the University of Lisbon, with a dissertation entitled "Factorial Analysis Models" under the supervision of Professor Fernando Nicolau, founding partner and co-responsible for the creation of the Association for Classification and Data Analysis (CLAD). Later, in Montpellier University (1983-1987) and under the supervision of Professor Yves Escoufier, he completed his PhD in Statistics and Data Analysis with the thesis Selection of Variables a Priori. During the period of obtaining these academic degrees, collaborations with the Universities of Minnesota (USA) and Upsala (Sweden) are highlighted as relevant in the area.

He began his academic career at the Faculty of Economics of the University of Porto, teaching between 1977 and 1991 subjects such as Statistics, Econometrics and Complements of Operational Research, having also been Director of the respective Data Analysis Laboratory. He has also taught, at the University Portucalense Infante D. Henrique (1988-2005), subjects such as Multivariate Data Analysis, Sampling Theory and Stochastic Processes, having been Director of the Degree in Statistics and Vice-Rector of this institution. He is currently, since 1994, Visiting Full Professor at NOVA IMS, School of Information Management and Data Science of the New University of Lisbon.

His non-academic professional activity is also intensely related to Statistics and Data Science. He was Regional Director of INE Porto, from 1989 to 2001, stimulating the production and dissemination of Portuguese official statistics and promoting cooperation with the Instituto Galego de Estatística. He coordinated the articulation of Statistics Portugal with the institutional world through the Statistical Council, of which he was later Vice-President from 2001 to 2003. He was National President of INE (Statistics Portugal) between 2001 and 2003, reinforcing INE's links with EUROSTAT and other official statistics producing organisations or Scientific Societies in issues underlying the national statistical production. He was President of the Association for Classification and Data Analysis from 2002 to 2004, promoting CLAD's articulation with other national and international scientific societies and the knowledge of Multivariate Data Analysis methods through the organisation of short courses in different practical contexts. It reinforced the CLAD-INE institutional relationship, cementing the link between the two institutions and establishing INE (CLAD's collective member) as a fundamental partner. He was part, in a regular and continuous way, of the Scientific Committee of the annual Conference on Data Classification and Analysis (JOCLAD), having presented the paper "Tipologia Sociodemográfica dos Concelhos da Região Norte: uma Análise Estatística Multivariada com base nos Censos de 1981 e 1991" at the first conference, held at the Faculty of Psychology and Education Sciences of the University of Lisbon, in 1993. He was part of the Jury of the first edition of the Fernando Nicolau Prize, 2021 edition.

Thematic Session
Bank of Portugal

18 July, 16:45 - 17:05

Complex made simple: Visualizing the business groups' database using Tom Sawyer

Diogo Silva[1], **Tiago Pinho Pereira**[2],
[1] Banco de Portugal, dfsilva@bportugal.pt
[2] Banco de Portugal, tppereira@bportugal.pt

The business groups' database of Banco de Portugal is the result of an algorithm that considers global group heads and relationships reported by Portuguese firms in Informação Empresarial Simplificada (Simplified Corporate Information, Portuguese acronym: IES). This database is used for developing business and external statistics, assessing credit risk, research and banking supervision purposes. Nowadays, in a very complex and globalised world, searching this database implies looking at very extensive, interconnected and sophisticated datasets and tables, which is especially challenging for less experienced users and practitioners. To take user efficiency to another level we have created an application based on the Tom Sawyer software that allows users to visualize business group structures in a very straightforward way. This software is very flexible since it enables users to select the desired degree of complexity, to change the design and the architecture of the visualization scheme and to highlight anomalies and errors in the data reported by companies, easing the quality control processes.

Keywords: Data visualization, business groups, quality control, Tom Sawyer

The business groups' database of Banco de Portugal contains tens of thousands of records by year and is used by several analysts for developing business and external statistics, assessing credit risk, research and banking supervision purposes. Its data source is IES. This is a mandatory survey for non-financial corporations in Portugal in which firms communicate their financial statements and additional tables detailing information. One of those further tables refers to related parties, which are entities that hold equity stakes in the declarant firm or entities in which the declarant firm have equity stakes. The global group head is also disclosed in related parties. Taking together all these ingredients, an algorithm uses all this information and builds the business groups' database, which corresponds to three datasets. One with resident entities and their group head. Other with non-resident entities and their group head. Finally, the last one comprises the relationships between entities, namely the identification of the participants and the investees, and the corresponding equity shares and voting rights.
To clearly observe the relationships between entities and across business groups we need a powerful data visualization tool. By reaching the records through a tabular view, users and

practitioners will miss a complete and rich picture that only a data visualization tool allows. Tom Sawyer software gives that unique experience to users and practitioners. This software reads the records from the three datasets and depicts the structure of each business group through a network graph, which combines nodes (the entities) and relationships (the equity participations), allowing a much more sophisticated and straightforward perspective. Moreover, it is highly flexible, since it enables zooming in, to get more information about two or three entities of the same business group, and zooming out, to know more about the relationships between two different business groups. It is also possible to change the layout of the network graph according to different needs. For example, the hierarchical layout is perfect for foreign direct investment analysis, since it allows for a top-down scheme, which makes it easier to identify the group head investing in firms from different countries. The circular layout is more useful when a business group is organized around different sectors of activity because it clusters firms around a common sub-group head (Figure 1).

(a) Hierarchical (b) Circular

Figure 1: Layouts

Visualizing data facilitates the quality control. For example, it is possible to select a doubtful relationship and identify the company that reported it, guiding the analyst towards the source of the issue. In addition, it is much easier to identify other problems such as entities (nodes) with no relationships (Figure 2). Analysts manually correct almost 5000 relationships every year, ensuring high-quality statistics and fulfilling the requirements of all users, which would not be possible without the help of Tom Sawyer.

(a) Before quality control (b) After quality control

Figure 2: The impact of data quality control

Disclaimer The analyses, opinions and findings of this paper represent the views of the authors, which are not necessarily those of the Banco de Portugal or the Eurosystem. Any errors and omissions are the sole responsibility of the authors.

18 July, 17:05 - 17:25

Live data – a game changer for official statistics

Elena Bucea[1], **Carlos Figueira**[2],
[1] Banco de Portugal, aebucea@bportugal.pt;
[2] Banco de Portugal, cfigueira@bportugal.pt

Visual and interactive dashboards are increasingly replacing plain numbers and tables. Visualization tools provide statisticians quick and easy-to-understand insights, allowing rapid analysis of patterns, outliers, as well as new phenomena, which in the past were hidden behind large and flat tables. Intuitive and friendly, these tools have made official statistics more attractive, allowing a deeper understanding of complex matters and the finding of answers to questions we did not know we had.

Keywords: Data visualization, Dashboards, Foreign direct investment, Official Statistics

---

Official statistics are crucial for a democratic society, serving the economy, the Government, and the public [1]. Statistics support the decision-making and the policy design process undertaken by Governments and international organizations and are used by the public and scholars for analysis purposes. The compilation of official statistics is, thus, a responsibility that goes far beyond producing numbers [3].

In an overloaded information society, official statistics face a major challenge – to be of practical utility. Users acknowledge the value and quality of statistics if they can access and interpret them easily. Thus, accessibility and flexibility are core features for on-demand analysis, meaning that statisticians have a crucial role on helping users in their interpretation without borders and limits. This requires moving from old communication outputs such as statistical yearbooks, full of tables, to innovative methods that build up statistical services tailored to users' needs.

Self-service visualization is one of the answers, making possible to untangle statistics' complexity, dynamics, and multidimensionality. New visualization tools have helped statisticians to enhance the understanding of raw data through a creative data exploration; graphics and interactive visualizations are much more appealing and comprehensible than plain numbers or tables. Analyzing data in a graphical format is significantly faster and thus enables compilers to address possible problems in a more timely manner. It is now possible to dig deeper into the data and find "answers to questions [we] did not know [we] had" (Ben Schneiderman). In a nutshell, these tools help to ensure and even increase data quality.

At Banco de Portugal, dynamic visuals have been playing a vital role in comprehending timely the most complex statistics and their different dimensions. Foreign Direct Investment (FDI) statistics cover all cross-border relations between enterprises from different

economies, representing long-term relationships between a direct investor and its direct investment enterprise. These statistics became increasingly complex over the years due to the globalization of multinational groups, the increasing deregulation of financial markets and permanent technological and financial innovations. In this context, FDI moved from one-to-one relationships to complex networks, increasing its multidimensionality. Given this, FDI statistics face the challenge of being too intricate, risking being misunderstood. Live data visuals have been a game changer for such statistics, by showing in an intuitive and friendly manner several aspects related to the cyclical nature of economic developments, the main ultimate and immediate investing countries, the currencies involved and the financial instruments used (see Figure 1).

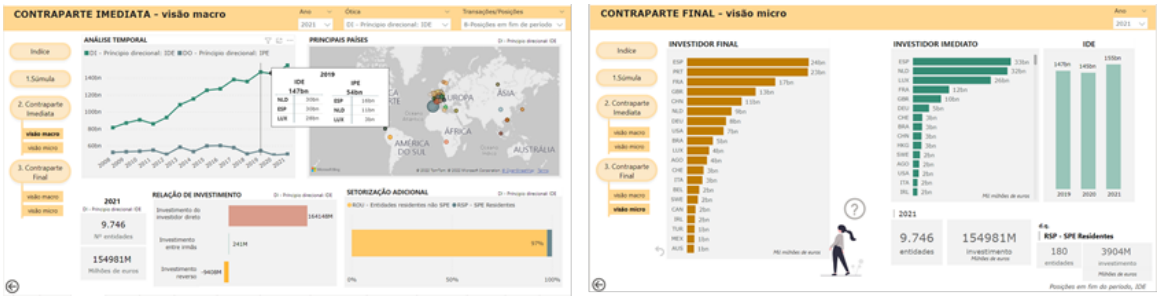

Figure 1: Dashboard on Foreign Direct Investment

Live dashboards ends the "one-analysis-fits-all", empowering compilers and users with personalized data filtering, clicking, and hovering over a graph to zoom in the data. The right visuals are fundamental for statisticians to compile high-quality data and to attract, engage and educate users, so that they make smarter, data-driven decisions. In this sense, visualization is a powerful tool to promote statistical literacy [2]. The notion of "click and see" has been broadening horizons, deepening the understanding of large complex datasets – abundant in official statistics – and capacitating users to be enthusiastic and informed analysts.

Disclaimer The analysis, opinions and findings of this paper represent the views of the author, which is not necessarily those of the Banco de Portugal or Eurosystem. Any errors or omissions are the sole responsibility of the author.

References

[1] United Nations. Fundamental principles of official statistics, 2014. *https://unstats.un.org/unsd/dnss/hb/E-fundamental%20principles_A4-WEB.pdf*

[2] Olivia Or. Data visualisation and its application in official statistics. In *Proceedings of the 59th ISI World Statistics Congress*, 2013.

[3] John Pullinger. Trust in official statistics and why it matters. *Statistical Journal of the IAOS*, 36(2):343–346, 2020.

18 July, 17:25 - 17:45

Give us data... and we give you back a wordless story

Carolina Rocha[1], **Mariana Oliveira**[2]
[1] Banco de Portugal, csrocha@bportugal.pt
[2] Banco de Portugal, moliveira@bportugal.pt

The relationship between enterprises, as reporting entities, and the central bank, as a statistical authority, is not confined to the legal obligations established between both parties. Enterprise and Sector Tables (EST) are a statistical product disclosed by Banco de Portugal, free of charge, and solely for the reporting enterprises. It allows them to compare their performance with the performance of their competitors. Despite being available since 2010, this product has changed throughout the years in order to provide a better service to the final user and consequently to promote better informed decision-making. EST are presented in the form of an interactive tool that organizes the data into themes such as, for instance, the funding structure, risk indicators and liquidity and cash, but can also be used just to get a quick and general overview of the enterprise. The goal is to communicate statistics based on three "storyteller commandments": provide complete data without overload, add value without too much complexity, and show a story using numbers without telling it with words.

Keywords: Data visualization, Enterprise and Sector Tables, Statistical communication

Long gone are the days when numbers published by statistical authorities, such as central banks, only intended to fulfill data needs of specific users, such as researchers or policy makers. In Banco de Portugal two phenomena are being observed: (i) a deeper users' segmentation, with the recognition of the public at large as also an important target group; and (ii) the commitment of engaging with each target group through offering of specific statistical content and products [2]. This paper focus on the Enterprise and Sector Tables (EST), a statistical output designed specifically for entrepreneurs, one of our main targets. EST was built over one of the biggest Banco de Portugal's database, which contains annual information on every single non-financial corporation operating in Portugal – around 400 thousand corporations - reported through the Simplified Corporate Information (Portuguese acronym: IES). IES is a mandatory report of the main financial statements (i.e., accounting data) and additional economic and financial information. IES contains about 3000 variables. This information is the main data source to compile aggregated economic and financial indicators on the non-financial corporations sector, such as, turnover, return

on equity or capital ratio. However, considering the increasing complexity of firms' activity, the exclusive offering of aggregated data would add limited value to them. Having this is mind, in 2018, the Banco de Portugal upgraded EST to a statistical product presented in the form of an online interactive dashboard that combines individual with aggregated data. Using EST, entrepreneurs can compare the performance of their company with the performance of every economic activity sector and with other companies with the same or different dimension. And it only costs them a few clicks. In order to guide the navigation and get the story, the dashboard is divided into themes, namely: highlights, activity and profitability, liquidity and cash, cash flows, funding structure, risk, balance sheet (structure), quartiles, and international comparison. Data is presented in the form of interactive graphics, with an intuitive navigation, simple language and with metadata to ease the interpretation. Each theme presents a set of predefined indicators with the aim of preserving the main message and avoid the overload of information.

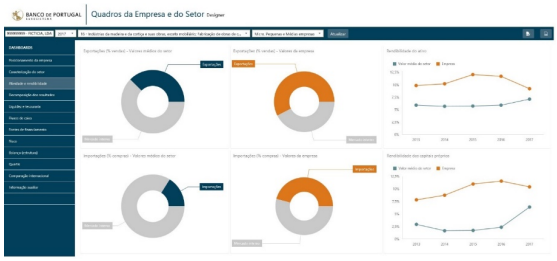

Figure 1: EST screenshot

For those who want a deeper analysis, it is possible to export a broader set of indicators to Excel (approximately 150 indicators per year), which facilitates the creation of their own dashboards [1].The interactive EST is an example of how to democratize the access to statistical information, turning easier even for less qualified entrepreneurs to understand their positioning in the market, and to make informed decisions. The recognition of the power of this service by entrepreneurs also triggers a virtuous circle of reporting higher quality data - through the IES - which will have higher quality information as a return - through EST. We call this, engagement.

Disclaimer The analyses, opinions and findings of this paper represent the views of the authors, which are not necessarily those of the Banco de Portugal or the Eurosystem. Any errors and omissions are the sole responsibility of the authors.

References

[1] Banco de Portugal. *Central Balance-Sheet Studies 36 Sector Tables and Enterprise Sector Tables*. 2019.

[2] R. Marques, L. Nunes, A. Colaço, and M. Oliveira. When reaching is no longer enough: 8 tips to engage with central banks' data users., journal=UNECE Conference of European Statisticians, year=2021.

Thematic Session
Statistics Portugal

18 July, 17:45 - 18:05

Classification of CPP - Application of a multilayer neural network

Almiro Moreira, Ana Carmona, M. Conceição Ferreira, David Santos, Rui Alves

Statistics Portugal

One of the most challenging problems in dealing with Census data is the classification of open answers, such as the job classification. The 2011 Portuguese Census data regarding individual jobs were limited to web answers collected through an open answer corresponding to more than 2.5 million cases. The rest of data were collected on paper and processed with OCR (Optical Character Recognition) and, due to their particular characteristics, were left aside in this work. The Standard Occupational Classification (SOC) or CPP classification (in Portuguese) is used as a standard to classify jobs in categories.
The CCP is the set of all professions existing in Portugal and their respective functional description, aggregated by professional groups. It is a fundamental instrument for statistics on occupations, both in terms of observation, analysis, consolidation of series and statistical technical coordination, and for statistical comparability at European and international level at all these common levels. The classification of occupations is relevant not only for the Census but also for other more regular statistical operations such as the Employment Survey (IE) or the Living Conditions and Income Survey (ICOR), for example.
In this work we use a 1-digit classification of the CPP of the 2011 Census (Large Group levels) in a multiclass classification problem (10 classes) by applying a multilayer neural network. Word Embeddings have been used, as a type of word representation that allows words with similar meaning to have a similar representation. Roughly speaking, word embedding, transforms text into numbers. Therefore, a technique like word embedding is used to map words or phrases from a vocabulary to a corresponding vector of real numbers. The algorithm used to learn word embedding was Embedding Layer.
Results show that that after evaluating the classes predicted in the test data, we find out that this model has an accuracy of 90%.

18 July, 18:05 - 18:25

Estimates on completed buildings for the Indicators System of Urban Operations – Exploring with ML methodologies

André Sousa, António Portugal, Inês Sá, Pedro Cunha, Sara Cerdeira

Statistics Portugal

Statistics Portugal publishes quarterly data on building permits for new constructions and on completed buildings. These data are based on the monthly permits issued by the 308 Municipalities across the country, under the scope of the Indicators System of Urban Operations (SIOU) – a system defined in 2002, based on the legislative changes resulting from the Legal Regime for Urbanization and Building. The goal behind SIOU is the gathering of data enabling us to follow the evolution of the building construction sector. It covers information on urban subdivision, land remodeling, building construction and demolition, completion of works and changes in use. The development of SIOU aimed at improving the reliability of information based on indicators and to obtain timely administrative data from the Municipal Councils, namely on the indicators relating to the completion of buildings. Given the administrative nature of the information, the data on building permits and completions are monthly updated and are subject to monthly and quarterly revisions. Delays in receiving some information from the municipalities, namely that concerning completion of works, are at the origin of the estimation performed by Statistics Portugal concerning the completion of buildings. The difference between the real deadline and the planned deadline for the conclusion of a building is estimated based on the planned deadline (i.e., the time elapsed between the permission to build and the effective conclusion of the building, as stated in the permit), based on a linear regression model, according to geography, the building characteristics, and its final use. Such estimation method aims at decreasing deviations resulting from revisions on the quarterly estimates.
The results obtained from the current methodology were compared with those obtained by the regression problem using machine learning algorithms. Additionally, algorithms for dealing with the estimation as a classification problem were explored. Regression models and decision trees were used (namely, linear or logistic models, and regression or classification trees, depending on the approach). In addition, boosting was applied to the decision trees algorithms in an attempt to reduce prediction errors. For each model, we defined fixed values for a set of parameters and let a tuning process guide the choice of values for the remaining ones. For the regression problem, the boosted regression tree resulted in the best predictions (RMSE = 364, MAE = 219, R2 = 0.341); for the classification problem, the best results were obtained from the classification tree (ROC AUC = 0.858). The

boosting did not improve the performance of the classification tree, but this was probably due to computing limitations that required a limited grid size for parameter searching in the tuning process.
The results suggest that, in this application, the classification models are better at distinguishing the cases of completion/non-completion of the building, than the regression models at estimating the difference between the real deadline and the planned deadline for the conclusion of a building. Alternatives for future research might include exploring a two-part model, where the classification and the regression approaches are combined, or taking a survival analysis approach to estimate the time to conclusion subject to censoring (e.g. in cases where no information on the building completion is received from the construction promoter).

References

[1] R. Tibshirani, T. Hastie, D. Witten, and G. James. *An Introduction to Statistical Learning with Applications in R.* 2021.

Author Index

Sponsors


CLAD
Associação Portuguesa de
Classificação e Análise de Dados


U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO


BANCO DE PORTUGAL
EUROSISTEMA


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL