DSSR 2020

10-11 DECEMBER 2020

Third international conference on
Data Science & Social Research

BOOK OF ABSTRACTS

Antonucci L., Kostiuk Y. eds. (2020). *Book of Abstracts*, Third international conference on Data Science & Social Research, Bari, 10-11 December 2020

ISBN 978-886629-051-3

PREFACE

This volume includes a selection of abstracts presented at the conference DS&SR 2020, held on December 10-11, 2020. It covers a wide variety of topics, in particular:

- new methodological developments to extract social knowledge from large-scale datasets;
- new social research about human behavior and society with large datasets, either mined from various sources (e.g. social media, communication systems) or created via controlled experiments;
- integrated systems to take advantage of new social data sources;
- Big Data quality issues, both as reformulation of traditional representativeness and validity and as emerging quality aspects such as access constraints, which may produce inequalities;
- Big Data for healthcare;
- text analytics and classification;
- sport analytics;
- environmental and spatial models;
- psychometrics.

The topics addressed are of wide relevance for data end social sciences.

Bari, Italy
December 2020

Laura Antonucci
Yana Kostiuk

COMMITTEES

SCIENTIFIC COMMITEE

Enrica Amaturo
Laura Antonucci
Biagio Aragona
Serena Arima
Tonio Di Battista
Gian Carlo Blangiardo
Giovanna Boccuzzo
Stefano Bronzini
Furio Camillo
Maurizio Carpita
Paola Cerchiello
Marcello Chiodi
Fabio Crescenzi
Corrado Crocetta
Francesco Domenico d'Ovidio
Giovanna Da Molin
Alessio Farcomeni
Maria Gabriella Grassia
Salvatore Ingrassia
Carlo Natale Lauro
Filomena Maggino
Paolo Mariani
Marina Marino
Antonello Maruotti
Antonella Massari
Stefania Mignani
Anna Paterno
Alessandra Petrucci
Domenico Piccolo
Alessio Pollice
Biagio Simonetti
Salvatore Strozza
Ernesto Toma
Nicola Torelli
Antonio Felice Uricchio
Rosanna Verde
Maurizio Vichi
Emma Zavarrone
Susanna Zaccarin

LOCAL ORGANIZING COMMITTEE

Corrado Crocetta
Laura Antonucci
Rossana Bray
Paolo Contini
Francesco Domenico d'Ovidio
Angela Maria D'Uggento
Francesca Falsetti
Fabio Manca
Claudia Marin
Antonella Massari
Paola Perchinunno
Alessio Pollice

CONTENTS

A DIGITAL ASSISTANT FOR SUPPORTING PATIENTS IN MONITORING ACTIVITIES

Marco Polignano, Marco de Gemmis and Giovanni Semeraro

Department of Computer Science, University of Bari, Aldo Moro, via E. Orabona 4, 70125, Italy
(e-mail: name.surname@uniba.it)

KEYWORDS: eHealth, chatbot, digital assistant, icd-10, clinical coding, medical reports

Abstract

The area of Computer Science that focuses on developing technologies to improve health, well-being, and healthcare is commonly known as eHealth. In this regard, we developed HealthAssistantBot, a Telegram-based conversational agent for supporting patients in their daily activities. In a simple way, by exploiting a natural language-based interaction, the system allows the user to create her health profile, describe her symptoms, search for doctors, or to remember a treatment to follow. We developed machine learning algorithms to deal with the patient necessity, such as the recommendation of the nearest doctor who can best treat her condition automatically identified by analyzing symptoms and medical records. The patient can use HealthAssistantBot for tracking everyday health parameters such as blood pressure, glycemia, weight, temperature, and heart rate. Moreover, the patient can use the platform as a repository to store all her clinical records in a secure way, including hospitalizations, medical checkups, and medical diagnosis. The user clinical conditions are automatically annotated using the ICD-10 glossary in order to make them interoperable and easy to understand for every doctor, everywhere in the world. We evaluated our HealthAssistantBot with both an offline and online evaluation obtaining encouraging results.

References

POLIGNANO, M., NARDUCCI, F., IOVINE, A., MUSTO, C., DE GEMMIS, M., & SEMERARO, G. 2020. *HealthAssistantBot: A Personal Health Assistant for the Italian Language*. IEEE Access, **8**, 107479-107497.

POLIGNANO, M., SURIANO, V., LOPS, P., DE GEMMIS, M. & SEMERARO, G. 2020. *A study of Machine Learning models for Clinical Coding of Medical Reports at CodiEsp 2020*. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS.

COMPUTATIONAL INTELLIGENCE FOR DIGITAL HEALTH: A BRIEF SUMMARY OF OUR RESEARCH WORK

Gabriella Casalino[0000−0003−0713−2260], Giovanna Castellano[0000−0002−6489−8628] and Gennaro Vessio [0000−0002−0883−2691]

[1] Department of Computer Science, University of Bari Aldo Moro, Bari – 70125, Italy
{gabriella.casalino,giovanna.castellano,gennaro.vessio}@uniba.it

Abstract

In the last few decades, a digitization process has involved various aspects of daily life, and the healthcare sector is one of the fields most heavily affected by this digital transformation. Artificial Intelligence, and in particular Computational Intelligence (CI) techniques, such as Neural Networks and Fuzzy Systems, have proven to be promising methods for extracting meaningful knowledge from medical data and for developing intelligent systems for faster diagnosis, im- proved monitoring and effective healthcare. CI-based systems can learn models from data that evolve as data changes, taking into account the uncertainty that characterizes health data and processes. Our group working at the Computational Intelligence Laboratory (CILab) of the Department of Computer Science, University of Bari, is currently carrying out scientific research on the application of CI techniques to Digital Health problems.

One activity concerns the creation of predictive models to support the early detection of episodes of Bipolar Disorder, a chronic mental illness characterized by the change of episodes that include healthy state, depression and mania or mixed states. We are investigating the effectiveness of applying semi-supervised learning to real-world data collected at the Department of Affective Disorders, Institute of Psychiatry and Neurology in Warsaw (Poland), through the interaction of patients with smartphones [1]. Another activity concerns the development of a monitoring system for the estimation of vital parameters based on the analysis of facial video frames [2]. The proposed system is based on a contact-less device (a videocamera integrated in a mirror), which is coupled with an intelligent module based on fuzzy logic rules to predict the level of risk of cardiovascular diseases. In collaboration with the local Institute for Biomedical Technologies of the Italian National Research Council, we are also working on the analysis of microRNA data from pediatric patients with Multiple Sclerosis [3]. This is a rare disease, the underlying mechanisms of which are still unknown. We want to support biology experts by not only providing some automated diagnostic tools, but also helping them find undiscovered patterns in the data.

References

[1] Casalino, G., Castellano, G., Galetta, F., Kaczmarek-Majer, K.: Dynamic incremental semi- supervised fuzzy c-means for bipolar disorder episode prediction. In: Discovery Science, 23rd In- ternational Conference, DS 2020, Thessaloniki, Greece, October 19–21, 2020, Proceedings, LNAI, vol. 12323, pp. 79–93. Springer (2020)
[2] Casalino, G., Castellano, G., Pasquadibisceglie, V., Zaza, G.: Contact-less real-time monitoring of cardiovascular risk using video imaging and fuzzy inference rules. Information **10**(1) (2019)
[3] Casalino, G., Vessio, G., Consiglio, A.: Evaluation of cognitive impairment in pediatric multiple scle- rosis with machine learning: An exploratory study of miRNA expressions. In: 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS). pp. 1–6. IEEE (2020)

TOWARDS AN USING OF BIG DATA IN HEALTHCARE: A LITERATURE REVIEW

Dicuonzo G.[1], Galeone G. [1], Massari A. [1]

[1] *Department of Economics, Management and Business Law, University of Bari*
(email:grazia.dicuonzo@uniba.it;graziana.galeone@uniba.it;antonella.massari@uniba.i)

KEYWORDS: 'Artificial Intelligence', 'Big Data Analytics', 'Healthcare'

The interest for new and more advanced technological solutions and, therefore, towards the use of Information and Communication Technologies (ICT) is paving the way for the spread of innovative and revolutionary applications in all business processes (Aceto et al., 2018). Even in healthcare organizations the demand for new and more advanced solutions from science and technology is becoming a solid reality. The Artificial Intelligence (IA) system applied to medical research has the potential to be able to diagnose, find vaccines or detect an epidemic and, therefore, and then move towards highly advanced e-Health (Tsikala Vafea et al., 2020). The pandemic emergency has led to extreme applications of artificial intelligence so that, through information resulting from the cross-reference of data from even heterogeneous sources, it is possible to draw deductions and derive correlations useful to predict clinical outcomes of patients, identify models that can lead to scientific findings that can pave the way for early diagnosis of Covid-19 infections while maximizing health care resources (Radanliev et al., 2020) and contributing to the containment of pandemic risk on national territory (Fusco et al., 2020). Effective and patient-centered care cannot disregard the acquisition, management and analysis of a huge volume and variety of health data and gleaning new insights from that analysis—which is part of what is known as Big Data (Bates et al., 2014)**.** The most relevant sources for the acquisition of large data in health care comes from medical recordings (Nambiar et al., 2013) as well as external data source. Additional data sources are increasingly available such us data derived from Internet use (social media) and smart application (Sadilek et al., 2012; Wicks et al., 2010). The potential acquisition and analysis of BDA requires a restructuring of the technological infrastructure and integrate traditional data analytical tools & techniques with an elaborate computational technology able to enhance and extract information useful for decision-making. Moreover it is important making the digital transformation of healthcare compatible with high standards of security and respect for data protection principles (Costa, 2014; Senthilkumar et al., 2018). This work aims to investigate the sustainability of Big Data use on the healthcare system in terms of improvement of services provided, changes in skills and abilities, efficiency of the organizational structure and operating costs.

References

ACETO, G., PERSICO, V., PESCAPÉ, A. 2018. The role of Information and Communication Technologies in healthcare: taxonomies, perspectives, and challenges. *Journal of Network and Computer Application*, **107**, 125–154.

BATES, D.W., SARIA, S., OHNO-MACHADO, L., SHAH, A., & ESCOBAR, G., 2014. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health* Affairs, **33**, 1123–1131.

COSTA, F.F., 2014. Big data in biomedicine. Drug Discovery Today **19**, 433–440.

FUSCO, A., DICUONZO, G., DELL'ATTI, V., & TATULLO, M., 2020. Blockchain in healthcare: Insights on COVID-19. *International Journal of Environmental Research and Public Health*, **17**, 1–12.

NAMBIAR, R., BHARDWAJ, R., SETHI, A., VARGHEESE, R., 2013. A look at challenges and opportunities of Big Data analytics in healthcare. *Proc. - 2013 IEEE International Conference on Big Data*, 17–22.

RADANLIEV, P., DE ROURE, D., WALTON, R., VAN KLEEK, M., MONTALVO, R.M., SANTOS, O., MADDOX, L.T., & CANNADY, S., 2020. COVID-19 what have we learned? The rise of social machines and connected devices in pandemic management following the concepts of predictive, preventive and personalized medicine. *EPMA Journal*, **11**, 311–332.

SADILEK, A., KAUTZ, H., SILENZIO, V., 2012. Modeling spread of disease from social interactions, in: Sixth AAAI International Conference on Weblogs and Social Media (ICWSM).

SENTHILKUMAR, S.A., BHARATENDARA, K.R., AMRUTA, A.M., G., A., CHANDRAKUMARMANGALAM, S., 2018. Big Data in Healthcare Management: A Review of Literature. *American Journal Theoretical Applied Business.* **4.**

TSIKALA VAFEA, M., ATALLA, E., GEORGAKAS, J., SHEHADEH, F., MYLONA, E.K., KALLIGEROS, & M., MYLONAKIS, E., 2020. Emerging Technologies for Use in the Study, Diagnosis, and Treatment of Patients with COVID-19. *Cellular and Molecular Bioengineering*, **13**, 249–257.

WICKS, P., MASSAGLI, M., FROST, J., BROWNSTEIN, C., OKUN, S., T., V., & BRANDLEY, R., J., H., 2010. Sharing health data for better outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, **12**.

KNOWLEDGE DISCOVERY BASED ON ARTIFICIAL INTELLIGENCE FROM MINING WEB SERVERS ACCESS LOG AND DATA VISUALIZATIONS

Amjad Zareen

Air University PAF Sector E-9, Islamabad, Pakistan
(E-MAIL: amjad.zareen@yahoo.com)

Abstract. Insufficient Logging and Monitoring significantly delays malicious activity or breach detection while carrying out incident response or digital forensics investigations. Web applications usage can be analyzed based on information available within access.log file in order to improve usability and security. Web Servers access.log contains significant amount of information that in turn can help to identify use and misuse cases. In this paper research has been presented to identify Web Servers behavior by producing meaningful visualizations which based on system generated access.log file. The followed technique has been implemented as an open source Python Jupyter notebook solution. The developed solution conforms to a newly proposed framework: Web Servers Access Log Visualizations – Attacks or Anomaly Detection using Applied Machine Learning. The developed solution has been tested and found functional with the size of access.log file greater than the available RAM of underlying machine.

KEYWORDS:: Web Server, Access Log, Analysis, Security, Python. Abstract

References

https://tools.ietf.org/html/rfc6872

Suneetha, K. R. and Dr. Krishnamoorthi, R. "Identifying User Behavior by Analyzing Web Server Access Log File", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

Pamutha, T., Chimphlee, S., Kimpon, C. and Sanguansat, P. "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns", International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol. 2, No. 2, June 2012.

Kharwar, A., Naik, C. and Desai, N. "A Complete Pre Processing Method for Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering, October 2013.

Punjani, M. and Gupta, V. "A Survey on Data Preprocessing in Web Usage Mining", IOSR Journal of Computer Engineering (IOSR-JCE), 2013.

Dr. Dhawan, S. and Lathwal, M. "Study of Preprocessing Methods in Web Server Logs", International Journal of Advanced Research in Computer Science and Software Engineering, 2013.

Verma, P. and Dr. Keswani, N. "Web Usage mining framework for Data Cleaning and IP address Identification", IJASCSE, 2014.

"TELEMIELOLAB" TELEMEDICINE PLATFORM ORIENTED ON DSS-AI

Alessandro Massaro[1,*], Fabio Manca[2], Angelo Galiano[1] , Giuseppe Calamita[2], Claudia Marin[2], and Angelo Vacca[2]

[1] Dyrecta Lab srl, Istituto di Ricerca, Conversano (BA), Italy.
(e-mail[*]: alessandro.massaro@dyrecta.com)

[2] Centro Interdisciplinare Ricerca Telemedicina, Università degli Studi di Bari Aldo Moro, Italy.

The "TeleMieloLab" is a pilot telemedicine platform developed within the framework of an Apulia region project oriented on the improvement of health assistance activities involving the telematic collaboration between Aziende Sanitarie Locali -ASL-, laboratory analyses, general practitioners, specialized doctors. The project is focused on the application on patients suffering from Multiple Myeloma. The platform is managed by a control room monitoring in real time the patients thus facilitating the pre-screening process through the use of a Decision Support System (DSS) integrating clinical algorithms with artificial intelligence (AI) ones predicting different parameters (MC, HB, Ca^{++}, IgC, IgA, IgM, Beta 2, FLC, etc.). The DSS is structured in three different alerting levels following a clinical flowchart. By using mobile app, the platform is able to enable homecare assistance processes. All patients are digitally traced by an identification number, and are automatically addressed about their analyses and therapies.

KEYWORDS: telemedicine, digital assistance, decision support system.

Acknowledgement

The work has been developed in the framework of the project: "Piattaforma di identificazione di nuovi marcatori prognostici nei pazienti con gammapatia monoclonale e di criteri di stratificazione terapeutica personalizzata 'TeleMieloLab'" (Bando INNOLABS POR Puglia FESR-FSE 2014- 2020).

References

MASSARO, A., GALIANO, A., SCARAFILE, D., FRASSANITO, A., MELACCIO A., SOLIMANDO, A., RIA, R., CALAMITA, G., BONOMO, M., VACCA, F., GALLONE, A., AND ATTIVISSIMO, F. 2020. Telemedicine DSS-AI multi level platform for monoclonal gammopathy assistance. *IEEE Proceeding of MeMeA 2020*, **1**, 1-

DSS INTEGRATED IN INFORMATION SYSTEMS ENABLING NEURAL NETWORK HOMECARE ASSISTANCE DECISION MAKING

Alessandro Massaro[1,*], Fabio Manca[2], Angelo Galiano[1], and Angelo Vacca[2]

[1] Dyrecta Lab srl, Istituto di Ricerca, Conversano (BA), Italy.
(e-mail[*]: alessandro.massaro@dyrecta.com)

[2] Centro Interdisciplinare Ricerca Telemedicina, Università degli Studi di Bari Aldo Moro, Italy.

Decision Support Systems (DSS) is an important tool for homecare assistance processes, also for companies working in homecare assistance services. The proposed DSS is based on neural networks predicting the patient health status. Another important DSS application is in in type I and II diabetes prediction, where Long Short-Term Memory (LSTM) neural networks provide a good accuracy. The neural network algorithms are integrated in the company information system in order to enable automatic alerting conditions and assistance procedures. The assistance control room receives in real time all patient data and plans the assistance service based on the processed alerting conditions.

KEYWORDS: telemedicine, neural networks, homecare assistance.

Acknowledgement

The work has been developed in the framework of the project: “Piattaforma B.I. intelligente di management risorse e di monitoraggio costi di assistenza sanitaria ‘Healthcare Assistance Platform: Management and Resources Allocation’” (developed for the company Assistenza Socio Sanitaria S.C.S.P.A.-Villa Puricelli, Piazza Puricelli 2, 21020 Bodio Lomnago (VA), Italy).

References

MASSARO, A., MARITATI, V., SAVINO, N., GALIANO, A., CONVERTINI, D., DE FONTE, E., AND DI MURO, M. 2018. A Study of a Health Resources Management Platform Integrating Neural Networks and DSS Telemedicine for Homecare Assistance. *Information*, **9**, 1-20.

MASSARO, A., MARITATI, V., GIANNONE, D., CONVERTINI, D., AND, GALIANO, A., 2019. LSTM DSS AUTOMATISM AND DATASET OPTIMIZATION FOR DIABETES PREDICTION, *Applied Sciences*, **9**, 1-22.

HEALTH WEARABLE DEVICES FOR PATIENT MONITORING AND ENABLING BIG DATA MULTI PARAMETRIC ANALYSIS BY ARTIFICIAL INTELLIGENCE

Alessandro Massaro[1,*], Fabio Manca[2], Angelo Galiano[1], and Angelo Vacca[2]

[1] Dyrecta Lab srl, Istituto di Ricerca, Conversano (BA), Italy.
(e-mail[*]: alessandro.massaro@dyrecta.com)

[2] Centro Interdisciplinare Ricerca Telemedicina, Università degli Studi di Bari Aldo Moro, Italy.

Health wearable devices are adopted to acquire daily patient physiological data. All data are collected into a Cassandra Big Data system and are processed by Artificial Intelligence (AI) algorithms such as Support Vector Machine (SVM) and Long Short Term Memory (LSTM). The outputs of the AI algorithms provide, with a good accuracy, the prediction of systemic vascular resistance, heart rate and blood pressure parameters. The study is focused on formulation of a four-dimension (4D) model (cube model updated during the time) mapping for each patient the health risk: the risk maps are based on the simultaneous analysis of multiple information including psychological score, physical activity, and environment pollution.

KEYWORDS: telemedicine, neural networks, big data, multi parametric model, wearable devices.

Acknowledgement

The work has been developed in the framework of the project: “Sistema di Supporto alle Decisioni con Intelligenza Artificiale e Modellizzazione Avanzata a più Dimensioni, per la Predizione ed il Raggiungimento dello stato di Equilibrio di Salute e per la Riabilitazione Ottimizzata mediante l’ Analisi di Dati Fisiologici da Bracciale e Big Data System ‘DSS/AI/BIG DATA WEARABLE HEALTH SENSOR” (developed for the company SANTA RITA GESTIONI Spa, Maglie (Le), Italy).

References

MASSARO, A., SELICATO, S., RICCI, GALIANO, A. AND RAMINELLI, S. 2020. Decisional Support System with Artificial Intelligence Oriented on Health Prediction Using a Wearable Device and Big Data. *IEEE Proceeding of MetroInd4.0&IoT 2020*, **1**, 718-723.

ARTIFICIAL NEURAL NETWORK APPLIED ON VOICE ANALYSIS FOR ONLINE REHABILITATION

Alessandro Massaro[1,*], Fabio Manca[2], Angelo Galiano[1], and Angelo Vacca[2]

[1] Dyrecta Lab srl, Istituto di Ricerca, Conversano (BA), Italy.
(e-mail*: alessandro.massaro@dyrecta.com)

[2] Centro Interdisciplinare Ricerca Telemedicina, Università degli Studi di Bari Aldo Moro, Italy.

The use of the internet for rehabilitation, is an actual approach to optimize the homecare assistance. A web platform is adopted to recognize the correct pronounced words and phrases in voice rehabilitation processes, by creating acoustic training model. The acoustic training model is implemented by a Long Shot Term Memory - LSTM- neural network algorithm, classifying the speech disorder and assigning a score for each test performed online. The platform is managed by a specialist which can select the exercise to perform thus analysing in real time the score assigned by the LSTM algorithm. The platform is able to trace the rehabilitation patterns and to suggest automatically the exercises to perform. The output graphical dashboards facilitate the clinical evaluations and reporting. The training dataset is structured as a "vocabulary" (dictionary) of the acoustic model able to classify the correct phonemes to pronounce. The innovative proposed approach of 'remote rehabilitation' provides a new concept of telemedicine based mainly on the goal to optimise human resources. A further improvement of the realized platform could be achieved by analysing the frequency response of the vocal signal.

KEYWORDS: telemedicine, neural networks, voice rehabilitation.

Acknowledgement

The work has been developed in the framework of the project: 'Programma computerizzato per il trattamento dei disturbi del linguaggio: 'VoiceAnalysis'' (developed for the company SANTA CHIARA srl, Muro Leccese (Le), Italy).

References

MASSARO, A., ET AL. 2020. Decisional Support System with Artificial Intelligence Oriented on Health Prediction Using a Wearable Device and Big Data. Accepted to *International Journal of Telemedicine and Clinical Practices*, (in press).

HEDONIC PRICES, TRANSPORT INFRASTRUCTURES AND GREEN AREAS: ECONOMIC EVALUATIONS AND REAL ESTATE MARKETS

Elisabetta Venezia[1]

[1] Department of Economics and Finance, University of Bari Aldo Moro
(e-mail: elisabetta.venezia@uniba.it)

Empirical studies in the field of property valuations are normally focused on structural and temporal attributes, in evolutionary terms, and only more recently have they focused attention, in quantitative terms, on those exogenous factors, of a spatial type, which always assume relevance growing. It is only in recent decades that there has been attention to these attributes that allows you to correct the estimates of value that would otherwise be distorted. Empirical evidence supports the hypothesis that greater accessibility to public transport and the presence of green areas could have a positive impact on real estate values. However, the capitalization of these advantages varies in different areas of study.
As indicated by Cordera et al. (2019), the weight of the different factors influencing property prices can be estimated using a technique known as hedonic regression. This technique was formalized by Rosen (1974). In this paper the methodology followed by Camagni and Capello (2003) is the starting point, nevertheless novels are included. They estimate a hedonic price function, indicated in the literature as a Box-Cox function, with the use of linear transformations. The analysis area for the evaluation of the hedonic model, here presented, is referred to a district located in the city center of Bari (an Italian city), north of the railway station, and with the presence of a green area. Therefore, it is suitable for assessing the effects on the differentials of price between cadastral and market values due to the presence of these two characteristics. 1,449 real estate units belonging to the category of economic housing are used to estimate the parameters of the model.
The results obtained are significant and the hypothesized relationship is correct.

KEYWORDS: Hedonic prices, Transport infrastructures, Green areas, Real estate market.

References

CAMAGNI, R, CAPELLO,R., 2003, Una valutazione ex ante di un grande progetto urbano attraverso la metodologia dei prezzi edonici, XXV Conferenza Italiana d Scienze Regionali.

CORDERA, R. ET AL.. 2019. The impact of accessibility by public transport on real estate values: A comparison between the cities of Rome and Santander, Transportation Research, Part A, **125**, pp. 308-319.

VENEZIA, E.,WILSON, P. ET AL..2013, Concorso internazionale di idee per il recupero delle aree ferroviarie del comune di Bari, awarded project.

A SYNTHETIC PENALIZED LOGITBOOST TO MODEL MORTGAGE LENDING WITH IMBALANCED DATA

Jessica Pesantez-Narvaez, Montserrat Guillen and Manuela Alcañiz

[1] Department of Econometrics, Riskcenter-IREA, Universitat de Barcelona, 08034 Barcelona, Spain; (e-mail: jessica.pesantez@ub.edu, mguillen@ub.edu, malcaniz@ub.edu)

Most classical econometric methods and tree boosting based algorithms tend to increase the prediction error with binary imbalanced data. We propose a Synthetic Penalized Logitboost based on weighting corrections. The procedure (i) improves the prediction performance under the phenomenon in question, (ii) allows interpretability since coefficients can get stabilized in the recursive procedure, and (iii) reduces the risk of overfitting. We consider a mortgage lending case study using publicly available data to illustrate our method. Results show that errors are smaller in many extreme prediction scores, outperforming a number of existing methods. Our interpretations are consistent with results obtained using a classic econometric model.

KEYWORDS: Imbalanced, boosting, interpretation, prediction, binary.

References

Barandela, R., Valdovinos, R. M., & Sánchez, J. S. (2003). New applications of ensembles of classifiers. *Pattern Analysis & Applications*, ***6***(3), 245-256.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth Int. Group, **37**(15), 237-251.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, ***16***, 321-357.

Dietterich, T. G., Domingos, P., Getoor, L., Muggleton, S., & Tadepalli, P. (2008). Structured machine learning: The next ten years. *Machine Learning*, ***73***(1), 3.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*, ***28***(2), 337-407.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189-1232.

Gomez-Verdejo, V., Arenas-Garcia, J., Ortega-Moral, M., & Figueiras-Vidal, A. R. (2005). Designing RBF classifiers for weighted boosting. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks* (Vol. 2, pp. 1057-1062). IEEE.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, **6**(5):429–449.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, ***9***(2), 137-163.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, ***30***(1), 25-36.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, ***5***(4), 221-232.

Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, ***409***, 17-26.

Longadge, R., Dongrre, S.S., & Malik, L. (2013). Class imbalance problem in data mining: Review. *International Journal of Computer Science and Network*, **2**(1): 83–87.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall/CRC.

Munnell, A. H., Tootell, G. M., Browne, L. E., & McEneaney, J. (1996). Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review*, 25-53.

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

Pesantez-Narvaez, J., & Guillen, M. (2020a). Penalized logistic regression to improve predictive capacity of rare events in surveys. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-11.

Pesantez-Narvaez J., & Guillen M. (2020b). Weighted Logistic Regression to Improve Predictive Performance in Insurance. *Advances in Intelligent Systems and Computing*, **894**, 22-34

Schapire, R. E., & Freund, Y. (2013). Boosting: Foundations and algorithms. *Kybernetes*.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, ***40***(1), 185-197.

Wang, S., & Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. In *2009 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 324-331). IEEE.

PREDICTING CONSUMPTION AND INCOME IN INDIAN HOUSEHOLD SURVEYS: A MACHINE LEARNING APPROACH

Nithin Raj K[1], Dr Althaf Shajahan[1]

[1] School of Management Studies, National Institute of Technology Calicut, India
(e-mail: nithinraj_m190021ms@nitc.ac.in, althaf@nitc.ac.in)

Our study aims at predicting the consumption of the household using asset measures, socio-economic indicators and household-level variables. It revolves around the theory of supervised machine learning, more precisely, regressions. We try to model the relationship between the consumption level as the dependant variable and various other independent variables. Machine learning works on the concept of regression where the model tries to find the best fit equation that closely explains the relationship. The model is trained using the training data and the accuracy is tested on testing dataset. We have selected three datasets for our study, viz, Indian Human Development Survey (IHDS) 2004 dataset, IHDS 2011 dataset and the All India Debt and Investment Survey (AIDIS) dataset. These nationally respresentative datasets contains the survey data on the household-level variables across India such as wealth measures, asset measures, social categories, whether urban or rural etc. They consists of 41554 observations for IHDS and 309026 observations in the case of AIDIS dataset. This is further resampled to 50000 samples to avoid overfitting problem. The majority of variables in IHDS datasets are categorical variables whereas most of the variables in AIDIS dataset are continuous variables. We really try to assess the scope of predicting the household consumption with the help of available independent variable. The machine learning algorithms we used include Ordinary Least Square regression (OLS), Ridge regression, LASSO regression, Random Forest regression, Elastic net regression and ensemble regressions which is based on bagging methods which utilizes the combination of different algorithms together. We find that the Random forest regression is the best fit model to predict consumption from the dataset with an accuracy of 63.27% in the testing dataset. OLS, ridge, LASSO and elastic net regressions shows an accuracy of 40.19%, 27.35%, 42.85% and 41.95% respectively. We also find that the prediction accuracy has significantly improved in the AIDIS dataset since it contains the data in the form of continuous variables whereas the variables in IHDS dataset were more categorical in nature. Asset measures, farming assets, livestock, social category, location (rural or urban) and household income levels seem to be predicting household consumption.

KEYWORDS: Consumption, survey data, machine learning, regression, prediction

References

MULLAINATHAN, SENDHIL, and JANN SPIESS. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31.2 (2017): 87-106.

ATHEY, SUSAN, and GUIDO W. IMBENS. "Machine learning methods that economists should know about." *Annual Review of Economics* 11 (2019): 685-725.

DYNAMIC CROWDING MAPS WITH MOBILE PHONE BIG DATA

Maurizio Carpita[1] and Rodolfo Metulini[2]

[1] Department of Economics and Management, University of Brescia,
(e-mail: maurizio.carpita@unibs.it)

[2] Department of Economics and Statistics, University of Salerno,
(e-mail: rmetulini@unisa.it)

In an environment subject to critical climate change, *smart cities* emergency management plans provide effective practices to decrease both citizen exposure and vulnerability. In this context, crowding maps are a valuable tool to enhance the flood risk management plans, and mobile phone big data help to explore spatio-temporal patterns and their uncertainties, with regard to land of interest and calendar period.

Recently, a multi-stage approach, based on Histogram of Oriented Gradients (HOG), Functional Data Analysis (FDA) and Model-Based Clustering (MBC) methods, has been used with spatio-temporal mobile phone big data along with administrative data to develop a dynamic indicator to estimate the number of citizen in an urban area (Metulini and Carpita, 2020).

In this talk, the use of this indicator to derive crowding maps is discussed. The proposed combined-methodology appears to be more reliable than standard crowdsourcing strategies, and has potentials to better address real-time rescues and reliefs supply. A test case is provided by a strongly urbanized area subject to frequent floodings located in the western outskirt of Brescia town in Northern Italy (Balistrocchi et al., 2020).

This Research Project is co-funded by Regione Lombardia, Call HUB Research & Innovation: "*Infrastrutture e servizi per la Mobilità Sostenibile e Resiliente* - MoSoRe@Unibs" ID 1180965 - POR FESR 2014-2020.

KEYWORDS: high-dimensional data, pattern recognition, decision support systems, smart city.

References

BALISTROCCHI, M., METULINI, R., CARPITA, M., & RANZI, R. 2020. Dynamic maps of people exposure to floods based on mobile phone data. *Natural Hazards and Earth System Sciences*, in press. DOI: 10.5194/nhess-2020-201.

METULINI, R., CARPITA, M. 2020. A spatio-temporal indicator for city users based on mobile phone signals and administrative data. *Social Indicators Research, online first*. DOI: 10.1007/s11205-020-02355-2.

I'VE SEEN THINGS YOU PEOPLE WOULDN'T BELIEVE: HOW DATA SCIENCE CAN HELP COACHING STAFF IN BASKETBALL

Manlio Migliorati[1]

[1] Department of Economics and Management, University of Brescia,
(e-mail: manlio.migliorati@unibs.it)

In this talk we will show some examples how data science can help basketball coach staff in tactically prepare and interpret a match, highlighting strengths to be pushed, or offering possible ways out when the situation seems compromised.
In particular we will show how classification and regression trees, thanks to their easy understanding, are a valuable tool for this purpose.
We will propose some examples where trees-based models are used to analyse match situations, to study basketball court on the base of shooting success and at last to build up more tailored match models via MOB.

KEYWORDS: data science, basketball, machine learning, trees

References

BREIMAN, L., & FRIEDMAN, J., & STONE, C. J., & OLSHEN, R. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC.

KUBATKO, J., & OLIVER, D., & PELTON, K., & ROSENBAUM, D. 2007. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, **3**(3):1-22.

MIGLIORATI, M. 2020. Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms. *Electronic Journal of Applied Statistical Analysis*, **13(**2), 454-473.

ZEILEIS A, & HOTHORN T, & HORNIK K 2008. "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 492–514.

ZUCCOLOTTO, P., & MANISERA, M., & SANDRI, M. 2017. Big data analytics for modelling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International journal of sports science & coaching*, **13**(4):569-589.

ZUCCOLOTTO, P., & SANDRI, & M. MANISERA, M., 2019. Spatial Performance Indicators and Graphs in Basketball. *Social Indicators Research*, Online First, 1-14

ZUCCOLOTTO, P. & MANISERA, M. 2020. *Basketball Data Science (with Applications in R)*. Chapman and Hall/CRC.

COMPLEX FEATURES IN REVIEW BOMBING

Venera Tomaselli[1], Giulio Giacomo Cantone[2] and Valeria Mazzeo[2]

[1] Department of Political and Social Sciences, University of Catania, *corresponding author*
(e-mail: venera.tomaselli@unict.it)

[2] Department of Physics and Astronomy "Ettore Majorana", University of Catania,
(e-mail: giulio.cantone@phd.unict.it, valeria.mazzeo@phd.unict.it)

'Review bombing' (RB) is a jargon referring to coordinated groups of people performing sabotage of a recommender system, e.g., submitting extreme scores. RB is conceptually alike to *fake news* or 'botting' on social media. The *Last of Us Part II* (TLOU2) is a survival horror videogame. It performed well under many rating metrics (e.g., Amazon's ratings from customers) since the publication day. However, in the platform *metacritic.com*, where it became also the most rated item, negative ratings outnumbered positive ones. By consensus, this was the widest case of RB on a single product. Contextual literature suggests a connection with political controversies over the videogame industry (i.e., #GamerGate). For each of approx. 65k different users having reviewed TLOU2 on *metacritic.com* in the first 40 days from the publication date, we scraped their review (i.e., text, score, and date) and other information about their history as reviewers. Many users submitted the review in the early days and significant differences in sentiment were observed among linguistic groups. Focusing on the largest English corpus of reviews (approx. 51k), we developed methods to identify relevant patterns, anomalies, and suspicious contents. We identified and quantified major topics (Politics, LGTBQ, boycott, etc.) under discussions. Through results, we argue for the existence of a quantifiable 'inverse effect' of RB, where negative reviews do fuel future positive reviews (and vice-versa) due to the inherent complexity of public opinion's dynamics.

KEYWORDS: review bombing, data mining, crowd-rating systems, *The Last of Us Part II*.

References

FERGUSON, C.J., & GLASGOW, B. 2020 Who are GamerGate? A descriptive study of individuals involved in the GamerGate controversy. *Psychology of Popular Media*. doi.org/10.1037/ppm0000280.

KASPER, P., KONCAR, P., SANTOS, T., & GUTL, C. 2019. On the Role of Score, Genre and Text in Helpfulness of Video Game Reviews on Metacritic. Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). doi:10.1109/snams.2019.8931866.

INTRODUCING SOME TESTS ON TRIAD CENSUS DISTRIBUTION WITH APPLICATION TO FOOTBALL

Lucio Palazzo[1], Riccardo Ievoli[2] and Giancarlo Ragozini[1]

[1] Department of Political Science, University of Naples "Federico II",
(e-mail: lucio.palazzo@unina.it, giragoz@unina.it)

[2] Department of Economics and Management, University of Ferrara,

(e-mail: riccardo.ievoli@unife.it)

Summary statistics of football matches such as final score, possession and percentage of completed passes are not satisfyingly informative about style of play seen on the pitch. In this sense, networks and graphs are able to quantify how teams are different from each others.

In this work we study the distribution of triad census, i.e. the distribution of local structures in networks that, basing on our results, is able to characterize passing networks of football teams. We describe the triadic structure and analyse its distribution under some specific probabilistic assumptions, introducing some tests to verify the presence of different triadic patterns in football data.

We firstly run an *omnibus* test against random structure, basing on U|MAN assumption, to assess whether observed triadic distribution deviates from randomness. Then we apply a Dirichlet-Multinomial test to recognize different triadic behaviours after choosing some reference patterns. The proposed tests are applied to a real dataset regarding 288 matches in the Group Stage of UEFA Champions League among three consecutive seasons.

KEYWORDS: count data, multivariate testing, passing network, sport performance, triad census

References

HOLLAND, P. W., & LEINHARDT, S. 1978. An omnibus test for social structure using triads. *Sociological Methods & Research*, **7**(2), 227-256.

LA ROSA, P. S., BROOKS, J. P., DEYCH, E et al. 2012. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one*, **7**(12), e52078.

USING EXPECTED GOALS IN A DOUBLE BINOMIAL MODEL FOR FOOTBALL OUTCOMES

Leonardo Egidi[1] and Nicola Torelli[1]

[1] Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste (e-mail: legidi@units.it, nicola.torelli@deams.units.it)

In modelling football outcomes, scores' data are regularly used for the estimation of the attack and the defence strength of each team. However, these teams' abilities are quite complex and are correlated with many quantities inherent to the game. Additional available information, relevant for their estimation and for a better analysis, are shots, both made and conceded, along with their geo-spatial structure, represented by the angle and the distance of each single shot in the football field. Following the motivations raised by the so-called expected goals (xG) models (Eggels et al., 2016) and extending a previous work about shots models (Egidi et al., 2018), we build a double binomial Bayesian model for the number of goals scored by the two competing teams, where the probability of scoring depends on shots covariates. We run our model on a large dataset consisting of the individual shots information for each match from three seasons of the Italian Serie A (2014-2015, 2015-2016, and 2016-2017) for a total of almost 27000 shots. Model checking through posterior predictive checks and model comparisons via leave-one-out cross validations are provided.

KEYWORDS: football modelling, shot information, double binomial model, Italian Serie A, model checking.

References

EGIDI L., PAULI F. & TORELLI, N. 2018 Are the shots predictive for the football results? *Book of Short Papers SIS 2018*, Pearson, ISBN-9788891910233.

EGGELS H, VAN ELK R. & PECHENIZKIY M. 2016 Expected goals in soccer: Explaining match results using predictive analytics. In: *The machine learning and data mining for sports analytics workshop*, p 16.

TRACKING IN FOOTBALL: FROM MOVEMENT COORDINATES TO THE CALCULATION OF DETAILED PARAMETERS AND REPORTING

Gianluca Rosso[1], Luca Malfatti[2]

[1] SIS Società Italiana di Statistica, quant4sport co-founder
(e-mail: gianluca.rosso@sis-statistica.org)

[2] Università di Torino Dipartimento di Management, quant4sport co-founder
(e-mail: luca.malfatti@unito.it)

The game analysis goes through tracking technologies. Detecting the relative positioning of the players and their coordinates is essential for carrying out analysis related to movement. There are wearable devices that use GPS signals but not all sports allow it, and they are not useful for tracking balls. For this reason, the optical tracking system is often used. Some sports have fewer complications than others: tennis for example. Others like football are difficult for the number of players, the duration of the game, the complexity of the movement. Quant4sport used professional tracking datasets to experiment with a visualization and reporting system useful for the ex-post analysis of the games: a system that only with the use of movement coordinates provides for the calculation of a series of parameters and their representation graphics. With an eye towards the future of artificial intelligence to use these parameters for predictions.

KEYWORDS: tracking, plotting, reporting, data analysis, key indicators.

FOOTBALL ANALYTICS: PERFORMANCE ANALYSIS DIFFERENTIATE BY ROLE

Mattia Cefis[1] and Maurizio Carpita[1]

[1] Department of Economics and Management, University of Brescia,

(e-mail: mattia.cefis@unibs.it, maurizio.carpita@unibs.it)

Nowadays, data science covers many areas of our life, and also sport applications. In this context, we focusing on football, and propose an overview about a project in the field of performance analysis (Carpita et al., 2019; Carpita and Golia, 2020) and prediction of football match results (Carpita et al., 2015).

The idea is to adopt a non-supervised approach, thanks some clustering around variables techniques (i.e. KPI: Key Performance Indicator), in order to create some composite index for each area of performance (e.g. technical-mental-physical), differentiate by role (Carpita et al., 2020). The final goal is to help coaches and scouting to take decisions and to evaluate impartially players performance.

In our presentation, we will submit an overview about the results of a preliminary analysis of our technical-dataset: data visualization and comparison between players KPI's performance, differentiate by role.

KEYWORDS: technical player performances, clustering around variables, composite indicators.

References

CARPITA, M., CIAVOLINO, E., & PASCA, P. 2020. Players' role-based performance composite indicators of soccer teams: A statistical perspective. *Social Indicators Research*, online first. DOI: 10.1007/s11205-020-02323-w.

CARPITA, M., GOLIA, S. 2020. Discovering associations between players' performance indicators and matches' results in the European Soccer Leagues. *Journal of Applied Statistics*, online first. DOI: 10.1080/02664763.2020.1772210.

CARPITA, M., CIAVOLINO, E., & PASCA, P. 2019. Exploring and modelling team performances of the Kaggle European Soccer Database. *Statistical Modelling*, 19(1). DOI: 10.1177/1471082X18810971.

CARPITA, M., SANDRI, M., SIMONETTO, A., & ZUCCOLOTTO, P. 2015. Discovering the Drivers of Football Match Outcomes with Data Mining. *Quality Technology & Quantitative Management*, 12(4). DOI: 10.1080/16843703.2015.11673436.

MACHINE LEARNING MODELS FOR THE EVALUATION OF STUDENTS' CAREERS

Rosa Arboretti[1], Riccardo Ceccato[2], Luca Pegoraro[2], Luigi Salmaso[2] and Cristina Tortora[3]

[1] Department of Civil, Environmental and Architectural Engineering, University of Padova, Padova (e-mail: rosa.arboretti@unipd.it)

[2] Department of Management and Engineering, University of Padova, Vicenza (e-mail: ceccato@gest.unipd.it, pegoraro@gest.unipd.it, luigi.salmaso@unipd.it)

[3] Department of Mathematics and Statistics, San José State University, San José, California (e-mail: cristina.tortora@sjsu.edu)

The evaluation of students' careers is fundamental for the organization of teaching and for the orientation and attraction of new students.

The School of Engineering at the University of Padova has collected data from different cohorts of students starting from the academic year 2013/14. Using demographic characteristics and information about a student's career, three main key performance indicators (KPIs) have been targeted: probability of achieving at least 55 University Credits at the end of the first year, dropping out at the end of the first year and graduating within 4 years.

Support Vector Machine [1], Random Forest [2] and Gradient Boosting Machine [3] were applied to the available data. In each application, the data set was randomly split into training and test sets and the performance of the model on the test set was measured, using the Adjusted Rand Index (ARI). For each classification problem, the model having the highest average ARI was selected. Finally, using the best performing models and appropriate importance measures, we identified the most relevant features having an impact on the probability of interest.

Among other results, this analysis allowed us to appreciate the efficiency of the entrance exam in evaluating students, given that the achieved grade is quite relevant when predicting the dropout probability. Moreover, satisfying eventual additional learning requirements (OFA) during the first year appears to be fundamental for the considered KPIs.

KEYWORDS: machine learning, university, academic career evaluation

References

[1] CORTES, C., & VAPNIK, V. 1995. Support-vector networks. *Machine learning*, 20(3), 273-297.

[2] BREIMAN, L. 2001. Random forests. *Machine learning*, 45(1), 5-32.

[3] FRIEDMAN, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

‘MINING’ THE SENTIMENT DURING COVID19 PANDEMIC

Albino Biafora[1], Angela Maria D’Uggento[1], Maria Gabriella Grassia[2] and Ernesto Toma[1]

[1] Department of Economics and Finance, University of Bari Aldo Moro,
(e-mail: angelamaria.duggento@uniba.it, ernesto.toma@uniba.it)

[2] Department of Social sciences, University of Naples,

(e-mail: mariagabriella.grassia@unina.it)

Influenced by the health emergency caused by the Covid-19 pandemic, many brands have changed their communication strategy, introducing more or less explicit references to principles of solidarity and fraternity in their TV commercial, to enhance trust and hope in Italian families during the lockdown. The traditional settings of the advertising format, focused on the product quality, have been moved to the background, in favour of the strengthening of the "brand image" by means of specific words, characters, hashtags and music, spreading empathetic messages to all those who need to regain hope and confidence, in a period of extreme emotional fragility. In this scenario, the paper aims to detect the emotions and the awareness raised by the brands during the lockdown period, compared to external periods, and to measure, through Text mining techniques, the sentiment of customers expressed on the social network Twitter. After a careful phase of extraction and pre-processing of tweets by means of Natural Language Processing algorithms, a dataset of 20,982 tweets was analysed, considering three different periods: pre-lockdown, during lockdown and post lockdown. The frequency analysis through the TF-IDF algorithm, firstly highlighted not only an increase in activity on the Twitter social network during the lockdown period of 3 times and half higher than the previous period, but, above all, a completely different use of the terms ‘Italia’, ‘Grazie’ and ‘Nuovo’, the most used in tweets, along with the hashtags '#iorestoacasa' and '#andràtuttobene’. The novelty in the advertising approach of the brands towards consumers, therefore, seems to have generated a certain sense of appreciation and gratitude, as well as a strong sense of belonging that is absent in the two other periods, in which it emerged that the TV commercial was commented on Twitter mostly for the brand's product or as an element of "disturbance" for other television events. The analyses of bi-grams and of co-occurrences highlighted that the commercials mostly drawing consumers’ interest were Lavazza and Barilla. In general, three reversals characterized the trend of the average daily “sentiment score”: a drop in “sentiment” just before the lockdown period, a strong resumption of positivity throughout it and a decline in the immediately following period. These results show how, by moving away from the traditional advertising approach, brands have managed, in order to enhance the amount of people talking about them, to reach the difficult goal of arousing an important positive mood in an extremely emotionally draining period; following, the trend has been reversed probably due to fatigue, and the first tastes of freedom.

KEYWORDS: ‘text mining’, ‘television spot’, ‘tweets’, ‘Covid19 pandemic’

UNIVERSITIES AND OPEN DATA, THE CHALLENGE HAS JUST BEGUN

Angela Maria D'Uggento[1], Rosa Ceglie[2] and Vincenzo Fiorentino[2]

[1] Department of Economics and Finance, University of Bari Aldo Moro,
(e-mail: angelamaria.duggento@uniba.it)

[2] Data Engineering Staff, University of Bari Aldo Moro,
(e-mail: rosa.ceglie@uniba.it;vincenzo.fiorentino@uniba.it)

Open data are spreading rapidly in public institutions (PI) as a means for promoting their effectiveness and efficiency and to enhance transparency, public participation, trust and cooperation. PI provide citizens with data and information regarding internal processes and policies, then data have to be easily accessible, understandable and usable. As to participation, shared decisional processes improve the quality of the political-administrative choices. On one side, PI gather proposals on citizens needs and, on the other side, build a cooperative network among themselves. Governance becomes a shared process which meets stakeholders' expectations. These are the pillars of Open Government, which has proved to be the most powerful way to face the recent economic crisis, turned into a confidence crisis towards Politics and institutions. The "openness" approach can spread these kinds of innovation: *i)institutional*, by rethinking the public intervention in terms of real usefulness for citizens/businesses; *ii)organizational*, based on transparency, organizational and individual evaluation, social responsibility and accounting; iii) *technological*, by creating a network of interconnected administrations, supported by modern technologies; *iv)cultural*, with the adoption of a participatory model, in which the decision-making process raises from the collaboration between institutions and individuals. In this scenario, data are crucial to provide citizens with the necessary knowledge tools to make their decisions or evaluate the impact of public ones. In a broader setting, the economic system can develop services that, based on public information, can benefit all the community. Public organizations collect a wide range of different data, which are particularly relevant as quantity and reliability. Universities received an important acceleration towards the adoption of 'innovative' IT procedures by the spread of the Covid19 pandemic, but, in addition to remote learning, they publish open data on institutional information: enrolled students, courses, departments, staff. Anyway, further information are more interesting for stakeholders: they would like to know if their investment in higher education will be rewarded by a successful university performance, by an adequate teaching or by effective possibilities of quickly entering the job market. Providing open data on these issues could be effective to compare the quality of universities or the potential value of a degree. Open data can lead to a further qualitative leap in the academic world, towards smart universities dealing with smart students, aware and motivated by a complete knowledge and participation in the institutions intended for them. Open data have a huge potential yet to be fullyexploited.

KEYWORDS: 'open data', 'open government', 'public university'

MACHINE LEARNING ALGORITHMS AND DESIGN OF EXPERIMENTS FOR PRODUCT INNOVATION

Rosa Arboretti[1], Riccardo Ceccato[2], Luca Pegoraro[2] and Luigi Salmaso[2]

[1] Department of Civil and Environmental Engineering, University of Padova, (e-mail: rosa.arboretti@unipd.it)

[2] Department of Management and Engineering, University of Padova, (e-mail: ceccato@gest.unipd.it, pegoraro@gest.unipd.it, luigi.salmaso@unipd.it)

Throughout the years statistics progressively gained importance in business and engineering contexts, to the point that today statisticians are often recognized as potentially leading innovation in the industrial environment (Hockman & Jensen, 2016). In the process of industrial innovation, a crucial part consists in the generation of knowledge about the phenomenon under investigation, thus traditional methodologies of industrial statistics such as Design of Experiments can be integrated with new tools coming from the field of Artificial Intelligence/Machine Learning in the aim of enabling the definition of new products with enhanced performances. In this work the authors present an application of machine learning methods on experimental data sampled according to a Box-Behnken design (Box & Behnken, 1960) in the process of development of a new chemical formulation which defines the properties of a commercial product. By using tools available in the literature, focus of the analysis is put on strategies which can be applied to illuminate the black-box of the machine learning models and provide a quantification of the uncertainty of predictions, points that are often under-emphasized in similar applications (Wager, Hastie, & Efron, 2014). The models developed can be included in a data-driven semi-automatic system that can facilitate the process of product innovation.

KEYWORDS: neural networks, random forests, experimental design

References

HOCKMAN, K.K., & JENSEN, W.A. 2016. Statisticians as innovation leaders. *Quality Engineering.*, **28**, 165-174.

BOX, G.E.P., & BEHNKEN, D.W. 1960. Some new three level designs for the study of quantitative variables. *Technometrics*., **2**, 455-475.

WAGER, S., HASTIE, T. & EFRON, B. 2014. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*., **15**, 1625-1651

"STRUCTURAL ANALYSIS" OF THE EU REGIONAL COMPETITIVENESS INDEX

Francesco D. d'Ovidio[1], Annamaria Fiore[2], and Rossana Mancarella[2]

[1] Department of Economics and Finance, University of Bari "Aldo Moro", Italy
(e-mail: francescodomenico.dovidio@uniba.it)

[2] ARTI-Puglia-Regional Agency for Technology and Innovation, Bari, Italy
(e-mail: a.fiore@arti.puglia.it, r.mancarella@arti.puglia.it)

ARTI-Puglia (Regional Agency for Technology and Innovation) recently published an instant report with many socio-economical comparisons between Puglia and the grouped rest of European regions, based on the EU Regional Competitiveness Index (https://ec.europa.eu/regional_policy/sources/docgener/work/2019_03_rci2019.pdf). The not only descriptive approach of the ARTI researchers provides numerous information on the characteristics of the RCI and some of its constituent elements.
The RCI is an indicator, created in 2010 by the Directorate General for Regional and Urban Policy of the European Commission, that is used to compare regional performances in terms of innovation and competitiveness, but also its social development. It takes into account about forty elementary indices at regional level (some of which, however, are not always available in some European regions, particularly in the French Overseas Territories), which identify eleven "pillars", then summarized in three macro-pillars. These three synthetic indicators are in turn summarized in the RCI, with different weightings according to the level of economic development of each observed region.
The work presented here aims to deepen the technical and applicative aspects of the RCI and its components, underlining its strengths and any critical issues in its structure. Through various multivariate techniques (starting with the Principal Component Analysis), the relevance and sensitivity of the numerous elementary indicators that are involved in the calculation of the RCI will also be verified.

KEYWORDS: Regional Competitiveness Index, Indicators, Structural analysis, PCA.

References

AGENZIA REGIONALE PER LA TECNOLOGIA E L'INNOVAZIONE, 2020. *Competitività ed innovazione: un confronto tra Puglia e regioni europee*. ARTI Instant Report, https://www.arti.puglia.it/sezione/scenari/pubblicazioni/instant-report.

ANNONI, P., & KOZOVSKA, K., 2010., *EU Regional Competitiveness Index 2010*, EUR 24346, Luxembourg: Publications Office of the European Union

KOZOVSKA, K, ANNONI, P., & DIJKSTRA, L, 2011. *A new regional competitiveness index: Theory, Methods and Findings*. European Union Regional Policy Working Papers, n. 02/2011.

MEASURING TOPIC COHERENCE THROUGH STATISTICALLY VALIDATED NETWORKS

Alessandro Albano[1] and Andrea Simonetti[1]

[1] Department of Economics, Business and Statistics, University of Palermo,
(e-mail: alessandro.albano@unipa.it, andrea.simonetti01@unipa.it)

Topic models arise from the need of understanding and exploring large text document collections and predicting their underlying structure. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has quickly become one of the most popular text modelling techniques. The idea is that documents are represented as random mixtures over latent topics, where a distribution over words characterizes each topic. Unfortunately, topic models give no guaranty on the interpretability of their outputs. The topics learned from texts may be characterized by a set of irrelevant or unchained words. Therefore, topic models require validation of the coherence of estimated topics. However, the automatic evaluation of the latent space of a topic model is a difficult task. Formerly, the most used metric for evaluating the quality of a topic model was the held-out likelihood. Still, the literature has shown that this method emphasizes complexity rather than interpretability. Although many procedures were recently proposed (Röder et al., 2015), the automatic evaluation of topic coherence remains an open research area. Our work aims to provide a new technique based on Statistically Validated Network (Tumminello et al., 2011). Our approach consists in representing each topic as a network of its most probable words. The presence of a link between each pair of words is assessed by statistically validating their co-occurrences in sentences against the null hypothesis of random co-occurrence. Thus, we propose a new coherence measure based on the structure of the statistically validated network. Furthermore, the new measure provides a ranking of topics and distinguishes high-quality from low-quality topics. The intuition is that the pairwise associations of words is strictly related to the semantic coherence and interpretability of a topic.

KEYWORDS: topic model, topic coherence, LDA, statistically validated networks.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 993-1022.

Röder, M., Both, A., & Hinneburg, A. 2015. Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15.*

Tumminello, M., Micciche, S., Lillo, F., Piilo, J., & Mantegna, R. N. 2011. Statistically validated networks in bipartite complex systems. *PloS one*, *6*(3), e17994.

MULTIVARIATE TEST FOR LARGE DATASETS AND SMALL SAMPLE SIZES

Stefano Bonnini[1]

[1] Department of Economics and Management, University of Ferrara,
(e-mail: stefano.bonnini@unife.it)

This work concerns an inferential methodology in order to deal with big datasets from the point of view of the number of response variables. In many testing problems, where the possible effect of a factor/treatment is evaluated under several points of view, in other words in the presence of a multivariate response, the classical methods based on the assumption of normality of the data cannot be applied, especially if the sample sizes are much lower than the number of variables (Pesarin and Salmaso, 2010).
The null hypothesis of the problem is that the treatment effect for all the considered variables is equal to zero. In other words it consists in the equality in distribution of the multivariate response for the two groups. The alternative hypothesis is the non-equality in distribution of the multivariate response for the two groups.
A suitable methodology for such a problem can be found within the family of the combined permutation tests (Pesarin and Salmaso, 2010; Bonnini et al. 2014). This methodology is nonparametric and does not require the dependence structure of variables to be explicitly defined, unlike the likelihood based methods, but it takes into account this structure in the null permutation distribution of the test statistic.
With this methodology, the property of finite sample consistency holds. When the number of response variables for which the alternative hypothesis is true diverges, the power of the test tends to one (Pesarin and Salmaso, 2010).
This solution is nonparametric, hence it is flexible and robust, with respect the departure from the assumption of normality.
We present a case study concerning a multivariate two-sample problem in a randomized controlled trial. It is a medical problem with a large number of response variables and very small sample sizes. No parametric solution is possible for such a problem.

KEYWORDS: permutation test, multivariate test, big data.

References

BONNINI, S., CORAIN, L., MAROZZI, M. & SALMASO, L. 2014. *Nonparametric Hypothesis Testing. Rank and Permutation Methods with Applications in R*. Wiley.

PESARIN, F. & SALMASO, L. 2010. *Permutation Tests for Complex Data*. Wiley.

A TOURIST SEGMENTATION BASED ON MOTIVATION, SATISFACTION AND PRIOR KNOWLEDGE WITH A SOCIO-ECONOMIC PROFILING: A CLUSTERING APPROACH WITH MIXED INFORMATION

Pierpaolo D'Urso[1], Livia De Giovanni[2], Marta Disegna[3], Riccardo Massari[1] and Vincenzina Vitale[1]

[1] Department of Social Sciences and Economics, Sapienza University of Roma, P.le Aldo Moro 5, 00185 Roma, Italy (e-mail: pierpaolo.durso@uniroma1.it, riccardo.massari@uniroma1.it, vincenzina.vitale@uniroma1.it)

[2] Department of Political Sciences, LUISS Guido Carli, Rome, Italy (e-mail: ldegiovanni@luiss.it)

[3] Accounting, Finance & Economics Department, The Business School, Bournemouth University, 89 Holdenhurst Road, Bournemouth, BH8 8EB, United Kingdom (e-mail: disegnam@bournemouth.ac.uk)

Cluster analysis is an exploratory multidimensional analysis of a complex dataset (Dolnicar, 2002; Ernst & Dolnicar, 2018; Dolnicar, 2019). Its main purpose is to discover homogeneous groups of units using a set of segmentation variables. In the last decades, the popularity of this tool is massively grown in the tourism filed but the review conducted in this study reveals that often the researchers are not aware of the characteristics and limitations of the clustering algorithms adopted. An important issue in the literature regards the adoption of the SPSS TwoStep clustering algorithm when mixed data are used as segmentation variables. In fact, it has been demonstrated that this algorithm is not suitable for this purpose and alternative algorithms should be adopted (Bacher et al., 2004). Therefore, the main purpose of this study is to describe, both theoretically and empirically, a novel clustering algorithm, recently published by D'Urso & Massari (2019), suitable to identify clusters of units based on mixed data. This clustering algorithm is so flexible that it works with any kind of data in input. Therefore, the second important contribution of this study is to discuss and present a suitable way to include the "Don't know" answer in the cluster analysis. As highlighted by Dolnicar (2013), the "Don't know" answer is frequently included in surveys but, as our review has revealed, this information has never been included in a cluster analysis conducted in the tourism field. Data collected at the GEOPARC Bletterback (South-Tyrol, Northern Italy), a UNESCO World Heritage site, have been analysed and discussed. Concluding, the

main issues related to cluster analysis are highlighted offering some suggestions and recommendations for future analysis.

KEYWORDS: fuzzy clustering, mixed data, don't know answers, visitors.

References

DOLNICAR, S. 2002. A review of data-driven market segmentation in tourism. *Journal of Travel & Tourism Marketing*, 12, 1–22.

ERNST, D., & DOLNICAR, S. 2018. How to avoid random market segmentation solutions. *Journal of Travel Research*, 57, 69–82.

DOLNICAR, S. 2019. Market segmentation analysis in tourism: a perspective paper. *Tourism Review*, ahead-of-print, ahead–of–print.

BACHER, J., WENZIG, K., & VOGLER, M. (2004). *SPSS TwoStep Cluster* - a first evaluation volume 2004-2.

D'URSO, P., & MASSARI, R. (2019). Fuzzy clustering of mixed data. *Information Sciences*, 505, 513–534.

DOLNICAR, S. (2013). Asking good survey questions. *Journal of Travel Research*, 52, 551–574.

RATIO OF FEMALES' INCOME OVER MALES' INCOME AND ITS DISTRIBUTION

Marcella Mazzoleni[1], Angiola Pollastri[2] and Vanda Tulli[2]

[1] Center on Economic, Social and Cooperation dynamics (CESC), University of Bergamo, (e-mail: marcella.mazzoleni@unibg.it)

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca (e-mail: angiola.pollastri@unimib.it, vanda.tulli@unimib.it)

In the analysis of gender gap, the difference between females' and males' income is one of the main topics as it is known that, even with a higher educational level, females earn less than males do.
To investigate this, we decide to estimate the distribution of the ratio of females' income over males' income using the methodology based on the distribution of the ratio of two Dagum with three parameters (Pollastri and Zambruno, 2010). The distribution of this ratio studied in two different situations can reveal the gender inequality concerning income in different groups or times. The parameters of the Dagum distribution, which fits very well to several economic variables' distributions, are estimated using maximum likelihood method.
We applied this methodology to the Bank of Italy Survey on Household Income and Wealth (SHIW) data analysing and comparing the deciles, the density functions and the cumulative distribution functions of the ratio of the females' income over males' income in different age classes, Italian areas, and years.
We decide to focus on this topic as this difference is decreasing in the recent years, but income parity has not yet been achieved.

KEYWORDS: Dagum distribution, ratio of two Dagum random variables, ratio of females' over males' income.

References

POLLASTRI A., & ZAMBRUNO G. (2010) Distribution of the ratio of two independent Dagum random variables. *Operations Research and Decisions*, **3** (20), 95-102.

DAGUM C. (1977). A new model of personal income distribution: specification and estimation. *Economie Appliquée*, **30** (3), 413-437.

THE MOST DEMANDING PASSAGES OF MATCH PLAY IN SERIE A SOCCER PLAYERS

Andrea Riboli [1,2]

[1] Department of Biomedical Sciences for Health, Università degli Studi di Milano

[2] Performance Department, Atalanta B.C., Bergamo, Italy

(e-mail: riboliandrea@outlook.com)

This study examined the most demanding passages of match play (MDP) and the effects of playing formation, ball-in-play (BiP) time and ball possession on the 1-min peak (1-min_{peak}) demand in elite soccer. During 18 official matches, 305 individual samples from 223 Italian Serie A soccer players were collected. MDP and 1-min_{peak} were calculated across playing position (central defenders, wide defenders, central midfielders, wide midfielders, wide forwards and forwards). Maximum relative ($m \cdot min^{-1}$) total distance (TD), high-speed running (HSR), very high-speed running (VHSR), sprint (SPR), acceleration/deceleration (Acc/Dec), estimated metabolic power (P_{met}) and high-metabolic load (HML) distance were calculated across different durations (1-5, 10, 90 min) using a rolling method. Additionally, 1-min_{peak} demand was compared across playing formation (3-4-1-2, 3-4-2-1, 3-5-2, 4-3-3, 4-4-2), BiP and ball/no-ball possession cycles. MDP showed *large* to *very- large* [effect-size (ES): 1.20/4.06] differences between 1-min_{peak} *vs* all durations for each parameter. In 1-min_{peak}, central midfielders and wide midfielders achieved greater TD and HSR (ES:0.43/1.13) while wide midfielders and wide forwards showed greater SPR and Acc/Dec (ES:0.30/1.15) than other positions. For VHSR, SPR and Acc/Dec 1-min_{peak} showed fourfold higher locomotor requirements than 90-min. 1-min_{peak} for Acc/Dec was highest in 4-3-3 for forwards, central and wide midfielders. 1-min_{Peak} was lower during peak BiP (BiP_{peak}) for HSR, VHSR and Acc/ Dec (ES: -2.57/-1.42). Comparing with vs without ball possession, BiP_{peak} was greater (ES: 0.06/1.48) in forwards and wide forwards and lower (ES:-2.12/-0.07) in central defenders and wide defenders. Positional differences in MDP, 1-min_{peak} and BiP_{peak} were observed. Soccer-specific drills should account for positional differences when conditioning players for the peak demands. This may help practitioners to bridge the training/match gap.

Key words: team sports; football; monitoring; performance; match load

Reference

Riboli, A., Semeria, M., Coratella, G., Esposito, F. 2021. *Effect of formation, ball in play and ball possession on peak demands in elite soccer.* **38**(2): 195-205.

SEARCHING FOR INCLUSIVENESS: A COMPARISON AMONG EUROPEAN TRAJECTORIES

Paolo Mariani[1], Andrea Marletta[1] and Mauro Mussini[1]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca
(e-mail: paolo.mariani@unimib.it, andrea.marletta@unimib.it, mauro.mussini1@unimib.it)

The Europe 2020 strategy is the EU program to support the growth and the occupation within 2020. About the economic growth, one of the three essential pursued priorities is the inclusiveness. In the year of its expiration, this contribution provides an exploratory analysis to verify the presence of an inclusive economic growth in some European countries. This study proposes an innovative approach based on trajectory analysis showing the dynamics of employment and well-being using a set of economic indicators for a 25-years period from 1995 to 2019. The obtained results have been compared to the economic models and social inclusion measures of 8 selected countries. The proposed methodology in this study is an approach for three-way data based on principal component analysis. In the two-dimension plan obtained by this method, it is possible to visualize a couple of coordinate for each year and each country. The union of these coordinates draws a single trajectory for each country defining the route across the period. Preliminary results have shown a lack of homogeneity in the economic growth and only in some cases, this dynamic could be targeted as inclusive.

KEYWORDS: inclusive growth, dynamic factorial analysis, Europe 2020, principal component analysis, time trajectory

References

BHALLA, S. 2007. Inclusive growth? Focus on employment. Social Scientist, 24-43.

D'URSO, P. 2000. Dissimilarity measures for time trajectories. Journal of the Italian Statistical Society, p. 53-83.

MARIANI, P., MARLETTA, A, MICHELANGELI, A. 2019. Inclusive growth in European countries: a cointegration analysis. 49th Scientific Meeting of the Italian Statistical Society.

ZHUANG, J., ALI, I. 2010. Poverty, inequality, and inclusive growth in Asia. Poverty, Inequality, and Inclusive Growth: Measurement, Policy Issues, and Country Studies, 1-32

TOPOLOGICAL METHODS FOR SOCIAL DATA ANALYSIS: AN OVERVIEW

Sara Scaramuccia[1,2], Roberto Fontana[1], Ulderico Fugacci[3], Francesco Vaccarino[1,2]

[1] Department of Mathematical Sciences "Giuseppe Luigi Lagrange", Politecnico di Torino, Torino, Italy (e-mail: sara.scaramuccia@polito.it, roberto.fontana@polito.it, francesco.vaccarino@polito.it)

[2] SmartData@PoliTO center on Big Data and Data Science, Torino, Italy

[3] Consiglio Nazionale delle Ricerche CNR, Istituto di Matematica Applicata e Tecnologie Informatiche, IMATI, Genova, Italy (e-mail: ulderico.fugacci@cnr.it)

In the context of social research, analysing data implies dealing more and more with large data sets, possibly with non-homogeneous coordinates, or abstract spaces. Even though a data set does not always come in the form of a shape, we can often build a shape on top of it by exploiting item relations, like in a network, or coordinate representations, like in an embedded point cloud. Topology describes, characterizes, and discriminates shapes by studying those properties which are preserved under continuous deformations, such as stretching and bending, but not tearing or gluing. Differently from geometrical measures, topological properties are coarse, independent from coordinate systems and stable to noise. The basic goal of topological data analysis [Carlsson, 2009][Patania et al., 2017] [Fugacci et al., 2016] is to infer knowledge from data by means of topology-based descriptors. We present an overview of the persistence pipeline as the main methodology in topological data. In this framework, the homology properties are tracked along a nested family of shapes to produce a barcode, that is, a data descriptor encoding the lifespan of connected components, loops, cavities and higher dimensional analogues along the nested family of shapes. We will focus on how long-lasting homology features are detected through the persistence approach and interpreted as the most relevant topological features in the family of shapes.
Are those topological features interpretable for social research? We discuss some persistence applications to complex networks by focusing on strength points and limitations which are specific to the social domain.

KEYWORDS: topological data analysis, networks, persistent homology, social sciences.

References

GUNNAR CARLSSON. "TOPOLOGY AND DATA." BULLETIN OF THE AMERICAN MATHEMATICAL SOCIETY 46.2 (2009): 255-308.

FUGACCI, ULDERICO, SARA SCARAMUCCIA, FEDERICO IURICICH, AND LEILA DE FLORIANI, "PERSISTENT HOMOLOGY: A STEP-BY-STEP INTRODUCTION FOR NEWCOMERS." STAG. (2016):1-10.

ALICE PATANIA, FRANCESCO VACCARINO, AND GIOVANNI PETRI. "TOPOLOGICAL ANALYSIS OF DATA." EPJ DATA SCIENCE 6 (2017): 1-6.

SOME EVIDENCE FROM DISTANCE LEARNING

Emma ZAVARRONE[1], Maria Gabriella GRASSIA[2], Rosanna CATALDO[2] and Rocco MAZZA[2]

[1] Department of Humanities, University IULM -Milan,
(e-mail: emma.zavarrone@iulm.it)

[2] Department of Social Sciences, University Federico II Naples (e-mail: mgrassia@unina.it; rcataldo@unina.it; rmazza@unina.it)

The Covid-19 pandemic has imposed an alternative scheme of learning than the traditional one. It is a paradox but an implicit benefit of pandemic was the propulsive thrust for adopting innovative learning methods. The distance learning was adopted with several differences among universities. Under this perspective, the teaching landscape has been deeply and quickly changed such as length and content of the lessons, students interaction, student reception. Therefore, old evaluation practices could not be appropriate and the mandatory survey on the quality and satisfaction of the students looses its meaning so another form of the students' evaluation has to be planned. At Iulm University, an approach based on textual analysis, was been adopted. Useful indicators, characterized by sentiment analysis score, have been developed for evaluating the distance learning. Under methodologic perspective the use of neural network models (Srinivasa-Desikan, 2018, Shayaa et al., 2018) on complete sentence for computing the sentiment score was preferred than the bag of words. Three opened questions related to the strengthens, weakness and suggestions aspects were submitted to Iulm students population (more of 7000 students), the response rate was 14% and the results highlighted a moderate positive sentiment for the distance learning but some shadows on lack both interaction students-professors and technical problems due to the low connectivity are captured by the proposed indicators.

KEYWORDS: textual analysis. machine learning, sentiment analysis

References

SHAYAA, S., JAAFAR, N. I., BAHRI, S., SULAIMAN, A., WAI, P. S., CHUNG, Y. W., ... & AL-GARADI, M. A. 2018. Sentiment analysis of big data: Methods, applications, and open challenges. *IEEE Access*, *6*, 37807-37827.

SRINIVASA-DESIKAN, B. 2018. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

CO.ME.T.A. – COVID-19 MEDIA TEXT ANALYTICS. A TEXTUAL DASHBOARD FOR POLICY-MAKERS

Emma ZAVARRONE[1], Maria Gabriella GRASSIA[2] and Rocco MAZZA[2]

[1] Department of Humanities, University IULM -Milan,
(e-mail: emma.zavarrone@iulm.it)

[2] Department of Social Sciences, University Federico II Naples (e-mail:
mgrassia@unina.it;rmazza@unina.it)

On Feb 11, 2020, WHO (World Health Organization) announced an official name for the syndrome coronavirus 2 (SARS-CoV-2), that is COVID-19. After a month the COVID-19 has been declared as pandemic. The focus of this paper is to trace how the mass media, particularly information on newspapers, have addressed the issues about the containment of contagion or the explanation of epidemiological evolution. Communication has an important role in the diffusion of behaviour and contagion, especially regarding the spreading of misinformation. It is important to monitor the information that the mass media and social platforms convey. Notable examples are Sharma et al. (2020), a Twitter based dashboard for analysing the COVID-19 misinformation, and Cinelli et al. (2020), who studied engagement and interest in the COVID-19 topic. This paper presents an interactive dashboard. CO.ME.T.A. based on textual methodologies of computational linguistic, textual analysis, classification, LDA, sentiment analysis, network textual analysis with a simple graphical interface. The sources of CO.ME.T.A. are the selected articles in newspaper on Covid-19 and its main purpose is to help the policy-makers in the comprehension of the events under quantitative perspective. The beta release of CO.ME.T.A. works on 45.673.657 terms realized on R and Python.

KEYWORDS: textual analysis. LDA, sentiment analysis

References

SHARMA, K., SEO, S., MENG, C., RAMBHATLA, S., DUA, A. AND LIU, Y. Coronavirus on Social Media: Analyzing Misinformation in Twitter Conversations. *arXiv preprint arXiv:2003.12309*, 2020.

CINELLI, M., QUATTROCIOCCHI, W., GALEAZZI, A., VALENSISE, C. M., BRUGNOLI, E., SCHMIDT, A. L. AND SCALA, A The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*, 2020.

DOCUMENT CLASSIFICATION VIA A MODEL-BASED APPROACH TO SPECTRAL CLUSTERING

Cinzia Di Nuzzo[1], Salvatore Ingrassia[1]

[1] Department of Economics and Business, University of Catania,
(e-mail: cinzia.dinuzzo@phd.unict.it, s.ingrassia@unict.it)

The interest in automatic procedures for handling textual data has grown exponentially, in particular for the text clustering algorithms that determine the natural clusters on the basis of a set of relevant words, see e.g. Aggarwal and Zhai (2012), Bécue-Bertaut (2018), among these, document clustering is used to group documents based on the same topic. In this context, the documents are numerically represented by vectors, see e.g. Salton et al. (1975), Melucci (2009), Salton and Buckley (1988), so the choice of a suitable similarity measure among documents and an appropriate clustering algorithm is important to identify the proximity of two feature vectors and the structure of the groups.

In this framework, we propose here a two-step model-based approach within the spectral clustering framework. Spectral clustering techniques have spread as valid competitors with respect to other clustering methods; they provide a data grouping using the eigenvectors of a matrix called the Laplacian matrix derived from a set of pairwise similarities among the units to be clustered, see e.g. von Luxburg (2007), Ng et al. (2002) and Meila (2015). Related approaches for text clustering have been previously considered in Cadot et al. (2018).

In the spectral clustering algorithm, the choice of the similarity between pairs of units plays a fundamental role; usually, similarities are computed according to a kernel function depending on some parameter ε called the radius. Our first contributions concern the proposal of a new kernel function for clustering text data via spectral clustering approaches and a data-driven criterion for the choice of the ε in the kernel function. Moreover, according to some theoretical results about the shapes of the clusters in the feature space, another proposal concerns a model-based approach to spectral clustering. In this framework, we compare different kinds of mixture models for grouping data in the eigenvector space. The performances of many different mixture models are compared based on real data sets.

KEYWORDS: spectral clustering, mixture models, document classification

References

AGGARWAL, CC. & ZHAI, C. 2012. A survey of text clustering algorithms. *Aggarwal CC, Zhai C (EDS) Mining Text Data, Springer, New York*, 77-128.
BÉCUE-BERTAUT, M. 2018. Textual Data Science with R. *CRC Press, Boca Raton.*

CADOT, M., &LELU, A., &ZITT, M. 2018. Benchmarking seventeen clustering methods on a text dataset. *Tech. rep., LORIA, hal-01532894v5.*

VON LUXBURG, U. 2007. A tutorial on spectral clustering. *Statistics and Computing.* **17**(4), 395-416.

MEILA M. 2015. Spectral clustering. *Hennig C, Meila M, Murtagh F, Rocci R (eds) Handbook of Cluster Analaysis, Chapman and Hall/CRC.*

MELUCCI, M. 2009. Vector-space model. *Encyclopedia of Database Systems, Springer.*

NG, A., & JORDAN, M., & WEISS, Y. 2002. On spectral clustering: Analysis and an algorithm. *Dietterich T, Becker S, Ghahramani Z (eds) Advances in neural information processing systems, MIT Press, Cambridge, MA,* **14**.

SALTON, G., & BUCKLEY, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* **24**(5), 513–523.

SALTON, G., & WONG, A. & YANG, C. *1975.* A vector space model for automatic index ing. *Communications of the ACM* **18**(11), 613–620.

AIR QUALITY IN LOMBARDY DURING THE COVID-19 BREAKDOWN

Paolo Maranzano[1] and Alessandro Fassò[2]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca,
(e-mail: paolo.maranzano@unimib.it)

[2] Department of Management, Information and Production Engineering, University of Bergamo
(e-mail: alessandro.fasso@unibg.it)

Newspapers, media, and even environmental agencies around Europe have reported that COVID-19 lockdown caused an extended environmental clean-up. Considering air quality, we focus on the Lombardy Region (Italy), which is at the same time the most populous region and the area most affected by COVID-19 in Italy. Lombardy is also one of the most polluted areas in the European Union. The main research hypotheses we pose concerns if and how first-wave restrictions imposed during spring 2020 have improved air quality in the region and if the improvements are similar throughout the territory. To answer these questions, we use data from January 2015 to mid-June 2020, provided by the regional environmental protection agency (ARPA Lombardia), for 74 monitoring stations. We used an autoregressive time series model with exogenous covariates (ARX) to assess the combined impact of meteorology, seasonality, trend, and lockdown on the NO_2 concentrations at each monitoring station. Model selection and inference are performed using a Least Absolute Shrinkage and Selection Operator (LASSO) algorithm. The statistical model confirms a generalised NO_2 reduction due to the lockdown. Compared to the observed average variations, the estimated lockdown impacts are mitigated by meteorology and natural trends, while showing strong and significant reductions in urban and industrialised areas.

KEYWORDS: air quality, COVID-9 lockdown, autoregressive exogenous covariates, LASSO algorithm, time series models.

References

FASSÒ, A. 2013. Statistical assessment of air quality interventions. *Stochastic environmental research and risk assessment*, 27(7),1651–1660. ISSN 1436-3240.

FINAZZI, F., & FASSÒ, A. 2020. The impact of the COVID-19 pandemic on Italian mobility, *Significance*, 17(3).

FASSÒ, A., & MARANZANO, P. 2020. Il cambiamento degli stili di vita e l'impatto della pandemia di COVID-19 sulla qualità dell'aria, *Statistica e Società*, Retrieved from http://www.rivista.sis-statistica.org/cms/?p=968

PARSIMONIOUS MATRIX NORMAL MIXTURES: AN APPLICATION TO UNIVERSITY STUDENTS INDICATORS

Salvatore D. Tomarchio[1*], Salvatore Ingrassia[1] and Volodymyr Melnykov[2]

[1] Department of Economics and Business, University of Catania
*Corresponding author: e-mail: daniele.tomarchio@unict.it

[2] Department of Information Systems, Statistics, and Management Science, University of Alabama

The analysis of the teaching efficiency plays a fundamental role for universities. This process can be measured by evaluating several aspects such as the successful completion of the studies by the students or the students/teacher's ratios. Here, these two areas are investigated for the University of Catania by using two matrix variate datasets. Such data structure allows the simultaneously consideration of several variables measured over different years. To detect possible underlying group structures, and for mitigating overparameterizations issues, parsimonious matrix normal mixtures are fitted to the datasets. The results show the existence of different groups in the data, each having distinctive characteristics, providing than useful information for the university managers.

KEYWORDS: matrix mixtures, parsimony, students' careers.

References

ANVUR. 2018. Rapporto biennale sullo stato del sistema universitario e della ricercar. Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca.

BELLOC, F., MARUOTTI, A., & PETRELLA, L. 2010. University drop-out: an Italian experience. *Higher education*., **60**, 127-138.

SARKAR, S., ZHU, X., MELNYKOV, V., & INGASSIA, S. 2019. On
parsimonious models for modeling matrix data. *Computational Statistics & Data Analysis.*, 106822.

SKIPTON, N., & COOPER, T. 2014. Business schools, student/teacher ratio and concerns for learning: Evidence from Canada. *Journal of Higher Education Theory & Practice*., **14**(3).

VIROLI, C. 2011. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*., **21**, 511-522.

THE GRAVITY PERCEPTION ON THE FISCAL FRAUDS IN ITALY: IS IT ONLY A QUESTION OF GEOGRAPHICAL AREA?

Paolo Aldrovandi[1], Ilaria Montaldi[1] and Mariangela Zenga[2]

[1] Department of Business and Law, University of Milano-Bicocca, Italy
(e-mail: paolo.aldrovandi@unimib.it, i.montaldi@campus.unimib.it)

[2] Department Statistics and Quantitative Methods, University of Milano-Bicocca, Italy
(e-mail: mariangela.zenga@unimib.it)

The study on fiscal fraud and tax evasion is a very complicated issue, due to the sensitive nature of the topic. In general, the respondents tend to underreport their practise on evasion when they are directly interviewed on a survey. In this paper we present an experimental online survey (n=350) where the respondents are interviewed on their perception on the fiscal frauds and on their propensity to evade taxes. The purpose is twofold: we firstly create an indicator on the gravity perception for fiscal frauds looking for socio-demographic differences. Secondly, we estimate the proportion of the tax evaders using the randomized response technique (Greenberg et al. 1969).

KEYWORDS: fiscal frauds, tax evasions, randomized response technique.

References

GREENBERG, B G, ABUL-ELA, AA, SIMMONS, W R & HORVITZ, D.G. 1969. The Unrelated Question Randomized Response Model: Theoretical Framework, *Journal of the American Statistical Association.*, **64**(326), 520-539.

THE FAKE NEWS DICTIONARY: AN OPPORTUNITY FOR MEDIA LITERACY

Rubaid Ashfaq[1], Zeba Nabi[2] and Rehan Irfan[2]

[1] Amity University, Noida,
(e-mail: rubaidashfaq@gmail.com)

[2] Lovely Professional University, Jalandhar

Fake news is not a new phenomenon, but the extent to which it can be reproduced on social networks is. When fake news is spoken about in various languages today, he realizes that phenomenon. The loss of centrality of the source and the possibility of "viralization" –another period term– often diminish the interest in the veracity of the news and the critical reading abilities to identify the false. To the extent that large proportions of the population are reported online, these issues have very direct political consequences, as seen in several recent events.

KEYWORDS: digital literacy, citizenship, fake, bubble filter, post-truth.

References

"OBAMA CHARGES AGAINST FAKE NEWS: 'IF WE CAN'T DISCRIMINATE BETWEEN SERIOUS ARGUMENTS AND PROPAGANDA, WE HAVE A PROBLEM '» IN LA VANGUARDIA, 11/18/2016.

"POPE WARNS THE MEDIA ABOUT THE 'SIN' OF SPREADING FALSE NEWS AND SLANDERING POLITICIANS" ON REUTERS, 12/7/2016.

SEE JEFFREY GOTTFRIED AND ELISA SHEARER: "NEWS USE ACROSS SOCIAL MEDIA PLATFORMS" AT PEW RESEARCH CENTER, <WWW.JOURNALISM.ORG/2016/05/26/NEWS-USE-ACROSS-SOCIAL-MEDIA-PLATFORMS- 2016/>, 26/05/2016

SEE NIC NEWMAN, DAVID AL LEVY, AND RASMUS KLEIS NIELSEN: REUTERS INSTITUTE DIGITAL NEWS REPORT 2015. TRACKING THE FUTURE OF NEWS 2, REUTERS INSTITUTE FOR THE STUDY OF JOURNALISM UNIVERSITY OXFORD, 2016.

TAURUS, BARCELONA, 2017

MICHELA DEL VICARIO, ALESSANDRO BESSI, FABIANA ZOLLO, FABIO PETRONI, ANTONIO SCALA, GUIDO CALDARELLA, H. EUGENE STANLEY, AND WALTER

QUATTROCIOCCHI: «THE SPREADING OF MISINFORMATION ONLINE »IN PNAS VOL. 113 N OR 3, 2017.

MARY MADDEN, AMANDA LENHART AND CLAIRE FONTAINE: "HOW YOUTH NAVIGATE THE NEWS LANDSCAPE", KNIGHT FOUNDATION, FEBRUARY 2017, AVAILABLE AT <HTTPS://KF-SITE PRODUCTION.S3.AMAZONAWS.COM/PUBLICATIONS / PDFS / 000/000/230 / ORIGINAL / YOUTH_NEWS.PDF>.

MICHAEL BARTHEL, AMY MITCHELL, AND JESSE HOLCOMB: «MANY AMERICANS BELIEVE FAKE NEWS IS SOWING CONFUSION »IN PEW RESEARCH CENTER, 12/15/2016.

IPSOS PUBLIC AFFAIRS: BUZZFEED FACEBOOK NEWS, 2017

AP - NORC CENTER / AMERICAN PRESS INSTITUTE: «'WHO SHARED IT?': HOW AMERICANS DECIDE WHAT NEWS TO TRUST ON SOCIAL MEDIA ", THE MEDIA INSIGHT PROJECT, MARCH 2017.

STANFORD HISTORY EDUCATION GROUP: «EVALUATING INFORMATION: THE CORNERSTONE OF CIVIC ONLINE REASONING », 2017

PHILIP N. HOWARD, GILLIAN BOLSOVER, BENCE KOLLANYI, SAMANTHA BRADSHAW, AND LISA-MARIA NEUDERT; «JUNK NEWS AND BOTS DURING THE US ELECTION: WHAT WERE MICHIGAN VOTERS SHARING OVERTWITTER? », COMPROP DATA MEMO 2017/1, 3/26/2017

FAKE SOCIAL MEDIA ACCOUNTS THAT ARE PROGRAMMED TO LIKE OR RETWEET A CERTAIN MESSAGE.

A. BESSI AND E. FERRARA: «SOCIAL BOTS DISTORT THE 2016 US PRESIDENTIAL ELECTION ONLINE DISCUSSION IN FIRST MONDAY VOL. 21 N OR 11, 11/2016.

L. GRAVES: DECIDING WHAT'S TRUE. THE RISE OF POLITICAL FACT-CHECKING IN AMERICAN JOURNALISM, COLUMBIA UNIVERSITY PRESS, NEW YORK, 2016

L. GRAVES AND FEDERICA CHERUBINI: "THE RISE OF FACT-CHECKING SITES IN EUROPE", REUTERS INSTITUTE FOR THE STUDY OF JOURNALISM / OXFORD UNIVERSITY, 2016.

AMONG THE TRADITIONAL MEDIA THAT HAVE DEVELOPED SOME INITIATIVE TO VERIFY THE VERACITY OF INFORMATION ARE LE MONDE (LES DÉCODEURS), LIBÉRATION (DÉSINTOX), FRANCE24 (LES OBSERVATEURS), THE WASHINGTON POST (FACT CHECKER), THE WALL STREET JOURNAL (BLUE FEED, RED FEED), CHANNEL 4 NEWS (FACTCHECK), THE GUARDIAN (REALITY CHECK AND BURST YOUR BUBBLE), BBC (REALITY CHECK), LA SEXTA (THE OBJECTIVE) AND EL PAÍS (FACTS)

J. KAHNE AND B. BOWYER: «EDUCATING FOR DEMOCRACY IN A PARTISAN AGE: CONFRONTING THE CHALLENGES OF MOTIVATED REASONING AND MISINFORMATION »IN AMERICAN EDUCATIONAL RESEARCH JOURNAL VOL. 54N OR 1, 2/2016.

P. MIHAILIDIS AND S. VIOTTY: «SPREADABLE SPECTACLE IN DIGITAL CULTURE. CIVIC EXPRESSION, FAKE NEWS, AND THE ROLE OF MEDIA LITERACIES IN 'POST-FACT' SOCIETY »IN AMERICAN BEHAVIORAL SCIENTIST, 3/27/2017

DAN GILLMOR: "FIX THE DEMAND SIDES OF NEWS TOO" IN NIEMANLAB, 2016.

SPATIAL CLUSTERING OF EUROPEAN NUTS 2 REGIONS BASED ON COVID DEATH RATES CHANGES

Andrea Bucci[1], Lara Fontanella[2], Luigi Ippoliti[1] and Pasquale Valentini[1]

[1] Department of Economics, Università degli Studi G. d'Annunzio Chieti-Pescara, (e-mail: andrea.bucci@unich.it, luigi.ippoliti@unich.it, pasquale.valentini@unich.it)

[2] Department of Juridical and Social Sciences, Università degli Studi G. d'Annunzio Chieti-Pescara, (e-mail: lara.fontanella@unich.it)

Since its outbreak in Wuhan city (China) in December 2019, the Corona Virus Disease (COVID-19) has spread widely throughout the world in a rapid and seemly uncontrolled way.
At the very beginning of this pandemic event, the most affected countries, such as Italy, France, Germany and Spain, implemented a series of large-scale interventions to control the epidemic. For example, the strictest control measures were applied in Italy with a complete lockdown of the population on March 11th.
Despite *Hsiang et al*. (2020) showed that this kind of interventions helped preventing or delaying around 530 million COVID-19 infections across several countries (China, France, Iran, Italy, South Korea, and USA), questions remain about the efficacy of these interventions. By using daily data on COVID-19 confirmed deaths (see also *Dergiades et al*. 2020) for 117 European NUTS-2 Regions, we propose a spatial clustering procedure of the time series to highlight areas of NUTS-2 regions showing similar government intervention and public health responses. Based on a set of features including, for example, the number and presence of changepoints *(Bai and Perron*, 2003), preliminary results suggest that, as expected, the detected phase changes differ among the NUTS-2 Regions with spatial patterns that change significantly between the first and last estimated changepoint.

KEYWORDS: COVID-19, changepoints, spatial cluster analysis

References

BAI, J., & PERRON, P. 2003. Computation and Analysis of Multiple Structural Change Models. *Journal of Applied Econometrics*, **18**, 1-22.

DERGIADES, T., & MILAS, C., & MOSSIALOS, E., & PANAGIOTIDIS, T. 2020. Effectiveness of Government Policies in Response to the COVID-19 Outbreak. Available at SSRN: https://ssrn.com/abstract=3602004 or http://dx.doi.org/10.2139/ssrn.3602004

HSIANG, S., & ALLEN, D., & ANNAN-PHAN, S., & BELL, K., & BOLLIGER, I., & CHONG, T., & DRUCKEMILLER, H., & HUANG, L. Y., & HULTGREN, A. & KRASOVICH, E., & LAU, P., & LEE, J., & ROLF, E., & TSENG, J., & WU, T. 2020. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*, **584**, 262-267.

A FRACTAL SAMPLING APPROACH FOR NETWORK ANALYSIS OF COVID-19 TWITTER DATA

R. Benedetti[2], E. Del Gobbo[1], S. Di Zio[1], L. Fontanella[1] and L. Ippoliti[2]

[1] Department of Legal and Social Sciences, Università degli Studi G. d'Annunzio Chieti-Pescara, (e-mail: emiliano.delgobbo@unich.it, s.dizio@unich.it, lara.fontanella@unich.it)

[2] Department of Economics, Università degli Studi G. d'Annunzio Chieti-Pescara, (e-mail: roberto.benedetti@unich.it, luigi.ippoliti@unich.it)

In some real-world problems, data are inherently represented as a graph. Such data, which are commonly referred to as network data, arise in the context of social networks, scientific collaboration, spreading of infectious diseases, company structures, etc. In such networks, nodes generally represent instances or particular objects of interest while edges, represented in an adjacency matrix A, provide information about their relations. By means of their nodes and edges, a graphical structure provides an effective, compact, flexible, and comprehensible representation of large-scale complex data (Aggarwal, 2011).

When dealing with large graphs, a wide variety of interesting applications of machine learning require the labelling of the nodes in the network. In fact, many of the major machine learning breakthroughs of the last decade have been catalysed by the release of labeled training datasets. Supervised learning approaches that use such datasets have thus increasingly become key building blocks of many classification tasks. For many real-world applications, however, large hand-labeled training sets do not exist, and are prohibitively expensive to create.

In this work, we are concerned with the problem of creating labelled training sets from an existing network. A common assumption in statistical machine learning is that the training set are independently and identically distributed (i.i.d.) according to a specified distribution D. That is, every instance (node) in the domain set is sampled according to D and then labelled according to an assumed "correct" labelling function, f. Here, to improve the quality of the training set, we focus on optimal spatial sampling designs strategies. The notion of optimal design is intuitive and corresponds to the objective of choosing n nodes in the network in an optimal fashion. There are numerous design criteria, such as A- or D- or G-optimality that have been extensively studied in a variety of contexts (see, for example, Dodge et al., 1988). In these and many other criteria, the major downside is that the optimality criterion depends on the model chosen as well as on the computational complexity of the chosen objective function. To avoid these problems, we introduce a space-

filling sampling strategy with a fractal-based objective function (Di Zio et al., 2004) evaluated over the patterns of points found in the adjacency matrix A. We shall show that the proposed procedure is very flexible and that it ensures a uniform coverage of the nodes of the network by also allowing the inclusion "spatial" constraints on possible patterns.

The procedure also easily adjusts to the case of semi-supervised learning where a set of seeds have to be selected. To showcase our method, we provide an example using Covid-19 twitter data to classify users opinions based on conspiracy theories.

KEYWORDS: supervised classification, network analysis, optimal spatial sampling, space filling curves, fractal dimension.

References

AGGARWAL C.C. 2011. *Social Network Data Analytics.* New York: Springer.

DI ZIO, S. & FONTANELLA, L. & IPPOLITI, L. 2004. Optimal spatial sampling schemes for environmental surveys. *Environmental and Ecological Statistics*, **11**, 397-414.

DODGE, Y. & FEDOROV, V. & WYNN, H. P. (1988). *Optimal Design and Analysis of Experiments*. Saint Louis: Elsevier Science Ltd

EXPLORING THE LINK BETWEEN AIR POLLUTION AND COVID-19 WITH ECOLOGICAL REGRESSION METHODS

Massimo Ventrucci[1] and Garritt L. Page[2]

[1] Department of Statistical Sciences, University of Bologna,
(e-mail: massimo.ventrucci@unibo.it)

[2] Department of Statistics, Brigham Young University, USA (e-mail: page@stat.byu.edu)

Recent studies have suggested that air pollution is associated with an increased risk of COVID-19 infection or death (Villeneuve and Goldberg, 2020). Most of these studies are based on ecological regression (ER) analyses of areal spatial data that is routinely collected by health and environmental agencies. ER analyses aim to estimate the relationship between a response variable, such as the area-level count of COVID-19 cases or deaths, and an exposure variable, such as area-level averages of fine particle matters (PM2.5). These models fall in the well-known class of generalized linear mixed models (GLMM) for spatial data. To provide reliable estimates of the effect of pollution on health, confounding variables and spatially structured random effects are often included in this framework, which allows adjustment for both measured and unmeasured confounders. For a review on the statistical challenges of ER methods see Bruno et al., (2017).

The main drawback of ER analyses is that inferences are made at the area-level, thus adjustment for individual-level confounding variables is not possible with these studies. For this reason, on the one hand ER is often seen as a valuable preliminary step before conducting individual-level studies that enable causal links to be investigated more accurately. On the other hand, ER can be very helpful to inform policy actions, often taken at the area-level, e.g. which populations/areas/social groups are to be prioritized in terms of resources allocation to prevent the spread of the pandemic.

We will discuss methodological issues involved in ER modelling of the association between COVID-19 mortality and exposure to air pollution. Our focus will be on the challenges introduced by using spatially structured random effects in a GLMM framework, including spatial confounding bias (Page et al., 2017). The models described will be illustrated on freely available USA county-level data (Wu et al., 2020).

KEYWORDS: 'areal data', 'Bayesian GLMM', 'spatial confounding', 'spatial modelling'.

References

BRUNO, F. ET AL. 2016. A survey on ecological regression for health hazard associated with air. *Spatial statistics*, **18**, 276-299.

PAGE, G.L. ET AL. 2017. Estimation and prediction in the presence of spatial confounding for spatial linear models. *Scandinavian Journal of Statistics*, **44** (3), 780-797.

VILLENEUVE, P.J. AND GOLDBERG, M.S. 2020. Methodological Consideration for Epidemiological Studies of Air Pollution and the SARS and COVID-19 Coronavirus Outbreaks. *Environmental Health Perspective*, **128** (9), CID: 095001.

WU, X. ET AL. 2020. Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Science Advances.*, **6**, eabd4049.

University of Bari *Aldo Moro*

ISBN 978-88-6629-051-3