

MECHANICAL ENGINEERING | PHYSICS |
PRESERVATION OF THE ARCHITECTURAL
HERITAGE | STRUCTURAL, SEISMIC
AND GEOTECHNICAL ENGINEERING |
URBAN PLANNING, DESIGN AND
POLICY | AEROSPACE ENGINEERING |
ARCHITECTURE, BUILT ENVIRONMENT
AND CONSTRUCTION ENGINEERING |
ARCHITECTURAL, URBAN AND INTERIOR
DESIGN | BIOENGINEERING | DATA ANALYTICS
AND DECISION SCIENCES | DESIGN |
ELECTRICAL ENGINEERING | ENERGY AND
NUCLEAR SCIENCE AND TECHNOLOGY |
ENVIRONMENTAL AND INFRASTRUCTURE
ENGINEERING | INDUSTRIAL CHEMISTRY AND
CHEMICAL ENGINEERING | INFORMATION
TECHNOLOGY | MANAGEMENT ENGINEERING
| MATERIALS ENGINEERING | MATHEMATICAL
MODELS AND METHODS IN ENGINEERING



Chair:
Prof. Piercesare Secchi

DOCTORAL PROGRAM IN DATA ANALYTICS AND DECISION SCIENCES

The Ph.D. program in Data Analytics and Decision Sciences (DADS) aims at training highly qualified senior data analysts and data managers capable of carrying out research at universities, international institutions, tech and financial companies, regulatory authorities, and other public bodies. The program stems from the cooperation between three departments: Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Dipartimento di Ingegneria Gestionale (DIG), Dipartimento di Matematica (DMAT), and the Center for Health Data Science Center of Human Technopole. It allows the enrolled students to work in a highly interdisciplinary environment with strong connections to international research centers and private companies. The program provides successful candidates with the opportunity to acquire a high degree of professional expertise in specific scientific and technological fields.

The program lasts three years: upon its successful completion and final exam, candidates will be awarded the title of Ph.D. in Data Analytics and Decision Sciences. The first year is devoted to the courses that build the broad competence and the solid interdisciplinary set of skills required by data analytics. The following two years focus on the development of the Doctoral thesis. Students must spend at least one semester in a research institution abroad, taking advantage of the network of international collaborations of the three departments involved in the program.

The program aims at breeding the next generation of data scientists who will tackle the challenges and the opportunities created by the increasing availability of the massive amount of data. These data scientists will be able to capture the relevant aspects of phenomena at play, develop adequate models, supervise the development of analytic pipelines, critically analyze the results, and support the technological transfer.

Data Analytics and Decision Sciences graduates are equipped with unique skills and advanced knowledge that open up career opportunities at universities, international research centers and institutions, R&D departments, regulatory authorities, financial institutions, tech companies, and other public bodies.

FACULTY BOARD

Prof. Secchi Piercesare
(Coordinator)

Prof. Azzone Giovanni

Prof. Caiani Enrico Gianluca

Prof. Ceri Stefano

Prof. Flori Andrea

Prof.ssa Ieva Francesca

Prof. Lanzi Pierluca

Prof. Matteucci Matteo

Prof.ssa Orsenigo Carlotta

Prof. Punzo Fabio

Prof. Roveri Manuel

Prof. Secchi Piercesare

Prof. Spagnolini Umberto

Prof.ssa Tanelli Mara

Prof. Tubaro Stefano

Prof.ssa Tumino Angela

Prof. Vantini Simone

ADVISORY BOARD

Ing. Elena Bottinelli
(Head of Innovation, digitalization del Gruppo San Donato)

Ing. Francesco Caio
(Dirigente pubblico e privato)

Prof. Emanuele Di Angelantonio
(Head of Health Data Science Centre, Fondazione Human Technopole)

ANALYZING ENERGY COMMODITIES AND THE EU ETS: STRATEGIES FOR MANAGING SHOCKS AND HEDGING IN THE LOW-CARBON TRANSITION

Mattia Chiappari - Supervisor: Andrea Flori

Tutor: Manuel Roveri

This thesis investigates the dynamic interplay between energy commodities and the European Union Emissions Trading System (EU ETS), offering new insights into how these markets respond to regulatory and economic shocks during the ongoing transition to a low-carbon economy. Through three distinct but interrelated studies, it addresses critical issues such as the propagation of market shocks, investor behavior leading to market anomalies, and the development of effective financial hedging strategies that align with environmental goals.

The first study examines the transmission of shocks across a comprehensive system of energy commodities, including fossil fuels, their derivatives, and European Union Allowances (EUAs). By applying advanced econometric techniques, such as spillover analysis in both time and frequency domains, this study provides a detailed view of how shocks originating from upstream (raw materials like oil and natural gas) or downstream (derivatives like gasoline and heating oil) energy markets propagate across the system. The analysis spans normal market conditions and periods of crisis, such as the COVID-19

pandemic and the Russia-Ukraine conflict. Results show that raw materials consistently act as dominant transmitters of shocks, especially during crisis periods, which amplifies market volatility. The study highlights the policy need for greater market oversight and more robust mechanisms to prevent the cascading effects of crises on energy markets, ensuring a more stable and secure energy supply during global disruptions. The second study focuses on herding behavior in the stock prices of firms regulated by the EU ETS. Herding occurs when investors mimic the behavior of others rather than relying on independent judgment, which can lead to inefficient pricing and increased market volatility. The study explores the relationship between herding behavior and spillovers in energy commodity markets, revealing that herding intensifies during periods of high market volatility and connectedness, such as regulatory changes or external shocks. This behavior distorts carbon prices, undermining the core function of the EU ETS, which is to reflect the true marginal cost of reducing greenhouse gas emissions. To address the financial implications of this

behavior, the study proposes a novel buy-sell trading strategy that incorporates both market volatility and spillover data to detect periods of heightened herding. The strategy identifies market conditions where herding is most likely and adjusts investment positions accordingly. The results of this strategy demonstrate superior risk-adjusted returns compared to traditional approaches such as equally weighted buy-and-hold portfolios and minimum connectedness portfolios. Specifically, the strategy was found to perform particularly well during periods of high market volatility, offering better protection against downside risks while taking advantage of the clustering effect of investor behavior. The practical application of this strategy suggests that combining market volatility with spillover information can yield profitable outcomes while also reducing the risks associated with irrational investor behavior. This insight is valuable for both asset managers and investors seeking to optimize their returns in volatile carbon markets. The third study proposes the development of a novel equity climate index based on the environmental performance of

firms regulated by the EU ETS. Unlike traditional ESG indices or Scope emissions, which often suffer from inconsistent and self-reported data, this index is grounded in verified emissions data, providing a more accurate measure of firms' carbon footprints. The research shows that this index offers superior hedging properties against a basket of financial assets, including stocks, bonds, commodities, and EUAs, particularly during the stricter regulatory phases of the EU ETS (Phases III and IV). However, the study notes that hedging with the climate index can be more costly compared to traditional financial instruments, although it provides better alignment with environmental objectives. This index offers a valuable tool for investors seeking to balance financial returns with sustainability goals, by investing in companies actively reducing their carbon emissions. In conclusion, this thesis provides a robust empirical analysis of energy commodities and the EU ETS, offering valuable insights for policymakers, investors, and market participants seeking to navigate the challenges of the low-carbon transition. By addressing key issues related

to market spillovers, herding behavior, and financial hedging, this research contributes to the development of more effective climate policies and sustainable investment strategies, ultimately supporting the global shift toward a greener economy.

STATISTICAL METHODS FOR HIERARCHICAL AND RECURRENT EVENT DATA IN LEARNING AND HEALTHCARE ANALYTICS

Alessandra Ragni - Supervisor: Anna Maria Paganoni

Analyzing data with grouping structures has become a prominent focus in statistical modeling in recent years. In real-world data analysis, complexities often arise from two primary sources: dependencies within hierarchical data structures and dependencies introduced by recurrent events. Hierarchical data involve units organized within nested groups or levels, such as students within schools or patients within hospitals. In the case of recurrent events, grouping is defined by repeated measurements or occurrences within the same statistical unit over time, creating dependencies – for instance, multiple hospital readmissions for a single patient or repeated student dropouts within the same degree program in higher education. These grouping structures may arise across any of the major types of data commonly analyzed in statistical research, including administrative, observational, and clinical trial data. The presence of grouping structures in real-world data, although challenging, necessitates careful consideration during the modelling process and motivates the use of tailored methods which are aware of the underlying structure. Failure to account for this aspect of the data can

lead to erroneous assumptions due to a misspecification of the dependency structure, ultimately resulting in conceptual mistakes in the formulation of the models and, thus, misleading results. Conversely, properly addressing grouping structures not only improves model accuracy but also yields more actionable insights, offering substantial support for policy-making decisions, which are often implemented at the group level. Within this context, the aim of this thesis is the development and application of novel models and methods suitable to handle hierarchical and recurrent event data, offering an in-depth understanding of the phenomena under study. This research focuses particularly on the fields of learning analytics and healthcare analytics, where grouped data often arise, with the goal of improving the evaluation of institutional effectiveness – such as in universities and hospitals – and the assessment of treatment or drug efficacy. Specifically, learning analytics utilizes data from surveys, performance records, and administrative datasets to identify student performance patterns and inform policy. Healthcare analytics, meanwhile, analyzes data from electronic

health records, clinical trials, and patient registries to improve patient outcomes, optimize clinical practices, and shape policies. Both fields frequently involve hierarchical and recurrent event data, making accurate statistical methods essential for enhancing overall system effectiveness. The statistical techniques employed and developed in this thesis build on mixed-effects and time-to-event models to enhance both the institutional effectiveness and the assessment of treatment efficacy, which are critical challenges in the learning and healthcare context. Upon these themes, this thesis is structured around two main parts, namely hierarchical data and recurrent event data, addressed through five distinct research lines, each presented in a dedicated chapter with a real-world case study. In Chapter 1, we address the critical task of identifying clusters within grouped data, a process essential for tailoring policies, particularly when the number of groups is large. We introduce an innovative approach within the framework of model-based clustering, leveraging linear mixed models with discrete random effects and exponential family distributions.

The proposed extension allows to identify latent clustering structures independent of fixed effects by assuming that random effects follow a discrete distribution with an a priori unknown number of support points. We apply our model to real-world data from the Programme for International Student Assessment (OECD), classifying countries based on their impact on the rates of low-achieving students in schools and providing valuable insights that can inform more effective educational policy formulation. In Chapter 2, we assess the impact of hybrid teaching – a combination of face-to-face and online learning – on student performance within engineering programs at Politecnico di Milano implemented in response to the Covid-19 pandemic. To achieve this, we employ multilevel propensity score techniques, introducing a novel pipeline in learning analytics which evaluates the impact of hybrid teaching by analyzing students' earned credits and grade point averages, while taking into account the grouping structure given by the specific engineering program in which students are enrolled. In Chapter 3, we leverage the same dataset to analyze time-to-dropout across different engineering programs, shifting our focus to survival outcomes. It is well-documented that Italian universities – including Politecnico di Milano – often experience high dropout rates which represent a waste of resources from an institutional perspective. This work aims to

provide insights that support the university in developing strategies to reduce dropout rates. Specifically, we extend the shared frailty model to accommodate time-varying frailty, allowing the frailty effect to change over time. We develop the TimeDepFrail R package based on the piecewise gamma frailty model originally proposed by Paik et al. (1994) and further refined by Wintrebret et al. (2004). Our extension includes a posterior frailty estimation, the computation of the conditional survival function and a significant reduction of the computational burden. Chapter 4 models student dropouts as recurrent events within university programs, nested within schools. Using Cox-based recurrent event models and functional data analysis, dropout intensity is mapped over time. A hierarchical extension of functional principal component analysis disentangles latent effects, improving dropout risk predictions and identifying critical dropout periods across programs. This novel approach refines dropout forecasting and enhances institutional decision-making processes. Chapter 5 tackles a significant challenge in clinical trials: assessing treatment effects when recurrent events terminate with death. When early deaths shorten the observation window, traditional methods may misrepresent treatment efficacy, emphasizing the need for a suitable estimator such as the expected number of events divided by the time alive within a specified time window, the

so-called while-alive strategy. This thesis proposes a highly efficient estimator specifically designed for randomized treatment settings. The estimator is applied to two real-world case studies, analyzing follow-up data from metastatic colorectal cancer patients in a phase III clinical trial and the HF-Action randomized controlled trial of chronic heart failure by O'Connor et al. (2011). Our goal is to assess significant treatment effects on the expected number of new lesion appearances in metastatic colorectal cancer patients and the number of hospitalizations in chronic heart failure patients over their time alive (up to t years). This ultimately demonstrates the practical advantages of the while-alive approach over state-of-the-art alternatives. These contributions advance existing methodologies and introduce novel statistical approaches for analyzing hierarchical and recurrent event data. The findings have broad implications for education and healthcare policy, improving both institutional effectiveness and patient or student outcomes.

ENHANCING THE ROLE OF REAL-WORLD DATA IN HEALTHCARE RESEARCH THROUGH ADVANCED STATISTICAL METHODS

Laura Savarè – Supervisor: Francesca Ieva

Co-Supervisor: Giovanni Corrao

In recent years, the expansion of large-scale healthcare utilization (HCU) databases has ushered in a transformative era in healthcare research, where empirical evidence increasingly informs clinical and policy decisions. This thesis underscores the pivotal role of real-world data (RWD) in generating empirical evidence essential for decision-making in clinical and policy settings, particularly within the Italian National Health System (NHS). While randomized clinical trials (RCTs) remain the gold standard for assessing treatment efficacy, they often fail to capture the full spectrum of real-world patient behaviors and the effectiveness of interventions in routine clinical practice. By leveraging advanced statistical methodologies, this research aims to unlock the full potential of RWD within the context of the Italian NHS, providing a more accurate representation of healthcare settings and addressing the limitations of controlled experimental studies. Through this approach, the study contributes to the generation of empirical evidence necessary to inform clinical practice and healthcare policy effectively. The core objective of this thesis is to enhance the methodological framework for analyzing

real-world data (RWD) to support evidence-based healthcare decisions. This research introduces and applies innovative statistical techniques, such as State Sequences Analysis (SSA) and Multichannel Sequence Analysis (MSA), alongside practical tools like healthcare dashboards. These methods enrich our understanding of treatment patterns and patient outcomes, setting a new standard for healthcare quality assessment and improvement. Furthermore, by incorporating a dynamic model with latent variables to capture the complexities of drug adherence in its multidimensional and evolving nature, this thesis provides a valuable framework for analyzing complex patient data. The Latent Markov Model (LMM) and the entire methodological pipeline (Fig.1) serves as a foundation for future studies addressing the challenges of modeling patient behavior in the presence of limited or incomplete data, such as in administrative databases. This novel approach not only tackles immediate issues related to comorbidity and polytherapy but also lays the groundwork for future advancements in healthcare analytics and patient behavior modeling. The application of these methodologies provides

novel insights into patient management, particularly for chronic conditions such as heart failure and severe mental health diseases. From a pharmacoepidemiological perspective, this thesis investigates methodologies to bridge the gap between RCTs and real-world applications, addressing challenges such as confounding, selection bias, and data heterogeneity. By assessing various pharmacoepidemiological techniques and validation strategies, the research evaluates their capacity to enhance the reliability of observational studies. The results demonstrate the value of integrating real-world evidence (RWE) into healthcare decision-making, reinforcing the role of observational data in complementing findings from clinical trials. A key focus of the study is therapeutic adherence, an essential determinant of treatment effectiveness. Through rigorous analytical approaches, this research evaluates adherence patterns and their implications for patient outcomes, providing evidence to optimize treatment strategies and improve patient engagement with prescribed therapies. The empirical applications of the proposed methodologies are illustrated through a series

of case studies covering different therapeutic areas. For instance, in the study of antihypertensive therapies, this research reveals the comparative effectiveness of monotherapy versus combination therapies, demonstrating how adherence influences cardiovascular outcomes. Similarly, in the case of T2DM, the findings emphasize the impact of adherence on clinical outcomes and healthcare costs, highlighting the role of newer pharmacological classes such as Sodium-Glucose Cotransporter-2 Inhibitors (SGLT2-I) and Glucagon-Like Peptide-1 Receptor Agonists (GLP-1 RA). The thesis also explores the use of data-driven modeling techniques in healthcare research, particularly in the development of predictive models for

COVID-19 risk stratification. These models contribute to early warning systems, improving resource allocation and patient management strategies during public health crises. A key contribution of this thesis is the integration of advanced statistical techniques with real-world data methodologies, enhancing the capacity of healthcare research to address complex questions about treatment effectiveness, cost-efficiency, and patient outcomes. By systematically applying these methodologies across multiple clinical domains, the research demonstrates the robustness of RWD in informing healthcare decisions. Furthermore, the findings advocate for the broader adoption of data-driven decision-making in healthcare, emphasizing the potential of

administrative health databases as a valuable resource for policy evaluation and healthcare planning. Ultimately, this work advances the field of healthcare analytics by showcasing the power of methodologically rigorous observational research. By overcoming traditional limitations associated with real-world data analysis, it lays the foundation for a more holistic, patient-centered approach to healthcare research. The integration of RWD-driven insights with clinical expertise paves the way for evidence-based healthcare interventions that are more responsive to patient needs and health system demands. The findings of this thesis contribute to the growing recognition of RWD as a cornerstone of modern healthcare research, supporting its role in shaping the future of precision medicine, health policy, and outcome-driven healthcare practices.

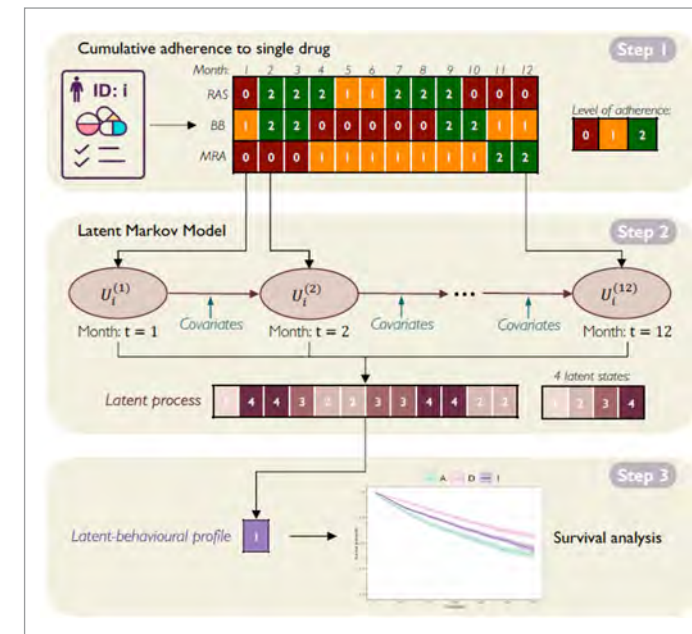


Fig. 1 - Schematic representation of the methodology used to develop a measure of adherence to polytherapy over time and study its association with patients' survival.