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

**PhD Yearbook | 2023**

# DOCTORAL PROGRAM IN DATA ANALYTICS AND DECISION SCIENCES

Chair:
Prof. Pier Luca Lanzi

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 at 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. Lanzi Pierluca (Coordinator) |
| Prof. Azzone Giovanni |
| Prof. Caiani Enrico Gianluca |
| Prof. Ceri Stefano |
| Prof. Di Angelantonio Emanuele |
| Prof. Flori Andrea |
| Prof.ssa Ieva Francesca |
| Prof. Mangiaracina Riccardo |
| 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. Vantini Simone |

# DATA INTEGRATION AND ETHICAL QUALITY: FUNDAMENTAL STEPS OF THE DATA ANALYSIS PIPELINE

**Fabio Azzalini** - Supervisor: Letizia Tanca

Data Science plays a very important role in the current society. In many scenarios, it allows to obtain insights that have a critical impact on our daily lives (e.g., precision medicine, fraud detection or autonomous vehicles), that otherwise would be impossible to achieve. Unfortunately, often the data sources used in data science applications are very heterogeneous and this prevents us from easily using them in data analytic tasks. In this context, before getting to the actual data modeling phase, it is necessary to apply a series of methods tso provide the data science algorithms with correct and reliable data. Specifically, often times the data comes from different sources that need to be integrated. Additionally, the data provided by the sources are often of poor quality, and can present ethical problems which, if not solved, would affect the final decisions of the prediction algorithms.

This thesis presents a collection of methods and tools to improve the quality of datasets and to prepare them for being used in data science tasks.

Data Integration is a fundamental task, and the role of a Data Integration System (DIS) is to address the modern challenges of a world where many different data sources of however structured information must interoperate, enabling users and applications to use them in a safe and consistent way. Therefore, data integration plays a very important role in determining the final performance of the analytics phase. Indeed, by combining different sources, it is even possible to create a resulting dataset of higher quality, often containing more and cleaner records with respect to the simple union of original data sources.

Typically, a data integration system operates according to the following three steps:

- Schema Alignment: when the data sources are structured, this phase has the purpose of aligning their different schemata and connect the attributes that have the same semantics.
- Entity Resolution: this phase has the purpose of finding, across the data sources, the records that match because they represent the same entities.
- Data Fusion: this phase has the purpose of applying the most appropriate techniques to merge the records that have been detected as matching in the previous phase.

Data Integration is a very broad and complex area of research, and in this thesis we focus on two of its sub-fields: blocking methods for entity resolution and data fusion.

The goal of Blocking methods is to partition the input dataset into blocks of similar records on which the matches are searched, thus greatly reducing the complexity of the problem. In some cases, with the large size of today data sources, blocking is even necessary to make the entity resolution step manageable. Yet, currently available methods still fail to satisfy the needs of many current heterogeneous sources.

To overcome the problems of the traditional methods we introduce LSH-Embeddings and Clust-Embeddings, two novel automatic blocking strategies that capture the semantic properties of data by means of Deep Learning techniques.
Both methods, in a first phase, exploit recent research on tuple and sentence embeddings to transform the database records into real-valued vectors. Sentence embeddigns are an evolution of word embeddings, techniques commonly used in the NLP field when analyzing textual data. Embeddings are used to transform word and sentences into numerical vectors that, when represented in a high dimensional space, place semantically related concepts close to each other in the space. Then, in a second phase, the two blocking methods employ different strategies to arrange the vectors inside the blocks: the first one adopts Approximate Nearest Neighborhood algorithms, while the other one uses dimensionality reduction techniques combined with clustering algorithms.
Adopting an unsupervised approach, we train our blocking models on an external, independent corpus, exploiting a "transfer learning" paradigm. Our choice is motivated by the fact that, in most data integration scenarios, no training data is actually available. We tested our systems on six popular datasets and compared their performance against five traditional blocking algorithms, demonstrating that our deep-learning-based blocking solutions outperform standard blocking algorithms, especially on textual and noisy data.

Despite the fact that Data Fusion is of great importance for constructing accurate datasets integrating sources of medium-to-low quality, only few techniques exist that are specifically designed to work with data items that may present multiple true values.

To address this issue, we have devised STORM, a novel algorithm for data fusion, designed to work well also in the multi-truth case. STORM extends the existing methods based on accuracy and correlation between sources in two respects.
First, it takes into account source authority; here, authoritative sources are defined as those having been copied by many other ones, assuming that when source administrators decide to copy data from other sources they choose the ones that they perceive as the most trustworthy. Second, the method envisages a value-clustering step that groups together the values recognized as being variants of the same real-world entity, thus reducing the possibility of making mistakes in the remaining part of the algorithm. Experimental results on three multi-truth real-world datasets show that STORM outperforms the eight best-performing state-of-the-art approaches.

On the side of the data integration problems we have just introduced, in this thesis we decided to consider Computer Ethics, a challenge that affects the quality of data sources and the predictions based on such data. When data science is used to build decision-making tools that impact the life of people, the problem of ethics becomes critically important, as a result we need to be sure that the data sources and the algorithms on which the analysis is based are fair and do not introduce bias in the decision process. As a consequence, in these particular applications of data analysis, data can be considered of good quality only if it conforms to high ethical standard.

Our research on Ethics has produced E-FAIR-DB, a novel solution that, based on constraints of a particular type, Functional Dependency, aims at restoring data equity by discovering and solving discrimination in datasets. The proposed solution is implemented as a pipeline, that firstly mines functional dependencies to detect and evaluate fairness and diversity in the input dataset, and then, based on these understandings and on the objective of the data analysis, mitigates data bias, with a particular attention to minimizing the number of modifications. Our tool can identify, through the mined dependencies, the attributes of the database that encompass discrimination (e.g. gender, ethnicity or religion); then, based on these dependencies, it determines the smallest amount of data that must be added and/or removed to mitigate such bias. We evaluated our proposal through theoretical considerations as well as with experiments on two real-world datasets. E-FAIR-DB can be utilized as a stand-alone tool or be used in a data integration pipeline to decide which sources have to be integrated, or in a data fusion algorithm to decrease the importance of unfair sources when determining the true values of each data item.

# A MACHINE LEARNING-BASED FRAMEWORK FOR AUTOMATIC AND INTERPRETABLE HEALTH AND USAGE MONITORING OF SAFETY-CRITICAL AIR AND GROUND VEHICLES

**Jessica Leoni** - Supervisor: Mara Tanelli

As technology has progressed, vehicles have become more advanced and complex, offering numerous benefits. However, these systems are often characterized by nonlinear dynamics that make it challenging to represent their behavior using first principle models. Indeed, despite the advantages provided, physics-based modeling performance is subjected to a trade-off between complexity and accuracy. It follows that accurate approximations may be too computationally intensive to be simulated, while less precise approximations can lead to misleading predictions.
This issue is particularly challenging for sophisticated vehicles composed of multiple critical subsystems, each with strongly nonlinear dynamics. Indeed, the lack of reliable models prevents to design robust diagnostic and prognostic monitoring platforms, which are critical for detecting and preventing system failures. In turn, this slows down the transition to a condition-based maintenance schedule, leading to costly vehicle downtime and decreased efficiency.

To address this issue, researchers have proposed increasingly sophisticated approaches to modeling complex vehicle systems. Among these, machine- and deep-learning techniques have emerged as particularly promising. Machine-learning uses input-output pairs to model a system's behavior and support decision-making, while deep-learning employs multi-layer neural networks providing even greater complexity. These techniques have proved highly effective in fault detection, thanks to their ability to describe complex dynamics and automatically identify anomalous conditions. However, this comes at the cost of reduced interpretability, as the complex structure of the classifier can obscure the logical steps underlying its predictions. As a result, machine- and deep-learning approaches are often referred to as "black-box" models, in contrast to the more transparent "white-box" physics-based models. It follows that, despite the advantages of machine- and deep-learning approaches, many end-users prefer traditional interpretable solutions, even if they are less accurate. Indeed, first principle models offer several benefits, including the ability to investigate the causes of detected damages, reduce false alarms, improve fault detection system performance, and enhance understanding of the monitored process.

Therefore, effort is still required to design machine- and deep-learning approaches that optimize the balance between accuracy and interpretability. One promising solution is offered by ensembles and mixtures, which leverage multiple shallow models to accurately capture complex dynamics. However, this approach tends to rely solely on black-box models and requires specification of which local model is appropriate for a given functioning condition - a specification that the end user may not be equipped to make. Various techniques have been proposed to address this challenge, including leveraging domain expertise, using a neural gating network, or employing clustering techniques to partition the data into different functioning modes and train a model on each one. However, these approaches tend to further reduce the interpretability of the model. This underscores the value of gray-box modeling, which combines white- and black-box models to optimize the strengths of both. Indeed, combining white- and black-box models in a gray-box approach can help overcome challenges related to data availability, system simplifications, and interpretability.
The first describes the nominal system behavior, while the latter is employed in case transients, nonlinearities, and anomalous conditions. However, gray-box modeling also requires knowledge of the boundaries that define the nominal dynamics to switch from white to black and vice versa adequately.

Effective design of diagnostic and prognostic systems requires addressing two critical challenges. The first challenge is the lack of interpretability of black-box models, which often results in end-users preferring less precise physics-based models. The second challenge is accurately estimating the validity region for each white- or black-box model included in the gray-box system. This thesis aims to provide practical solutions to these issues by presenting a set of methods to develop data-driven and interpretable health and usage monitoring platforms that rely on advanced machine and deep learning techniques. The proposed frameworks are designed to be interpretable, and extensive case studies have been conducted for air and ground vehicle applications. The thesis proposes black-box health and usage monitoring frameworks to address transmission vibration monitoring, engine monitoring, and flight condition recognition for helicopters. Moreover, functionalities such as road quality recognition, driving style characterization, rider mass estimation, and second passenger detection have been implemented for electric scooters, and early accident detection for motorcycles. These frameworks have the potential to revolutionize the field of diagnostics and prognostics by enabling the development of more accurate and interpretable systems, thereby enhancing passenger safety and reducing operational costs.

Furthermore, this thesis also presents a novel methodological contribution defined as the Automatic Physics-Informed Mixture of Experts (API-MoE), which combines data-driven and physics-based models to provide an optimal description of a system's behavior. The API-MoE not only learns the local models during training but also their reliability based on the system's operating conditions. As reported in the Figure below, the approach involves two stages of learning: inferring the local model parameters and identifying the model that relates a feature vector to the reliability of the local models.

The API-MoE combines multiple interpretable black-box methods and physics-based formulations of the system's dynamics, following the guidelines presented in the case studies' frameworks. This flexible and scalable method can manage any number of white- and black-box models. The use of lasso regularization helps achieve sparsity by considering the minimum number of relevant local models, improving interpretability and reducing computational intensity. The API-MoE is highly effective in fault detection, enabling the development of effective prognostics and diagnostics platforms, improving passenger safety and reducing operational costs. The methodology integrates domain experts' prior knowledge with insights gained from analyzing a large amount of historical data on the system's variables.

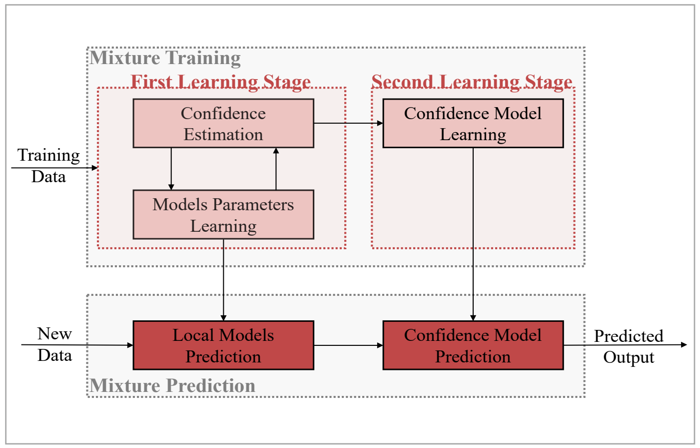


Fig. 1

# THE IMPACT OF REGIONAL DEVELOPMENT POLICIES ON SOCIO-ECONOMIC DEVELOPMENT AND SUSTAINABILITY TRANSITION

**Francesco Scotti** - Supervisor: Fabio Pammolli

This dissertation investigates the impact of the European Cohesion Policy, representing a cornerstone among place-based policies implemented at European Union (EU) level. Overall, my thesis is composed by three research papers that aim to assess the impact of the EU Cohesion Policy across different priority investment areas. Through the integration of heterogeneous sources of data and the application of state of the art statistical analyses, this work contributes to show the effect of EU Cohesion policy at the intersection of the social, economic, environmental and innovation dimensions.

More specifically, Paper I analyses the immediate and medium-long term impact of SCFs expenditures across different sectors. Furthermore, it investigates the relevance of spatial and technological spillovers generated by these funds. Through the application of a generalized propensity score matching model with a continuous treatment and spatial panel models based on geographical and technological proximity matrices over the time frame 2007-2014, I show that the energy, research and development and transportation sectors significantly contribute to economic growth with persistent effects in a medium-long term perspective. Conversely, the environment sector displays a slightly negative immediate impact on GDP per capita, complemented by positive effects at different time lags. Furthermore, I show that the transportation sector, that accounts for the largest portion of the EU budget over the analysed time frame, produces the highest spillovers, thus contributing to the local growth also of close regions from a geographical and technological perspective. Finally, I highlight that the highest spillovers are generated by provinces in Belgium, while at national level, Belgium, Germany, the Netherlands and UK produce the highest aggregate spillovers.

Paper II analyses some limitations of the current criterion implemented by the European Commission (EC) to allocate the largest portion of SCFs, devoted to territories with a GDP per capita below the 75% of EU average. In particular, I identify two different groups of regions potentially penalized in terms of received SCFs as a consequence of the EU enlargement occurred in 2004. The former (Not treated again (NTA) regions) encompasses EU-15 regions that were not classified in the group of less developed regions both in the programming period 2007-2013 and 2014- 2020, but that would have been likely to receive the less developed status in absence of the EU enlargement. The latter (Lost treatment (LT) regions) includes NUTS-2 that lost the status of less developed regions for the timeframe 2014- 2020, but that would have been likely to receive the less developed status in absence of the EU enlargement (based on the pre-enlargement benchmark represented by the 75% of EU-15 average GDP per capita). The application of state of the art Synthetic Control Methods and Difference in Differences approaches at different levels of geographical scale provides evidence of a strong economic penalization of NTA regions subject to a significantly lower GDP per capita growth between −10.5% and −5.7%. Conversely, LT regions did not experience significantly lower economic growth. This might be due to the fact that such regions received a "safety net", a financial cushion that guaranteed them to obtain at least two-thirds of the budget received in the previous programming period.

Paper III studies the impact of different combinations of policy instruments adopted at EU level in order to foster the sustainability transition. I consider SCFs and H2020 funds as main examples of technology push policies, since they represent the most relevant instruments adopted by the EC to tackle climate change. Furthermore, I take into account the European Union Emissions Trading System (EU ETS) as an example of demand pull policy instrument, since it covers more than 45% of overall GHG emissions at EU level. Through the application of a dynamic two ways fixed effects (TWFE) panel event study, I show that the EU ETS has a positive impact on innovation performances of firms in the manufacturing sector, contributing to raise patents applications between 0.4% and 0.5% during the period 2013-2020 (Phase III). Conversely, the effect generated by this demand pull policy instrument is not significant over the previous two phases (2005-2007 and 2008-2012) and when I consider either the energy sector, or firms across all economic sectors. Such empirical evidence suggests that the effectiveness of demand pull policy instruments depends both on policy stringency and on market conditions that may affect the extent to which firms can avoid to fully internalize the carbon price in their investment decisions. On the other hand, when I consider the policy mix, SCFs and H2020 allocated to environmental innovation projects generate additional benefits with respect to the EU ETS policy only for the energy sector. In particular, SCFs and H2020 increase patents applications in the energy sector between 1.4% and 2.1%, with a reduction of CO2 emissions in the range 0.7%-0.8%. This result points to the fact that in case firms can adopt alternative strategies to reduce the impact of the carbon price in their business activities, technology push instruments represent effective complementary measures with respect to demand pull policies, improving the environmental performances of the underlying firms. Finally, I analyse the mechanism explaining why a properly designed policy mix based on the combination of demand pull and technology push instruments may generate additional environmental innovation at firm level. In this direction, I show that SCFs and H2020 do not crowd out investors, but catalyze capital, activating additional research and development activities, contributing to raise patents applications and decrease CO2 emissions.

Overall, this thesis contributes to the debate on the effectiveness of the EU Cohesion Policy, providing robust empirical evidence with respect to several significant aspects of this regulatory framework. Although my strongest effort to implement methodologically grounded research, some limitations still affect my work and may open future research opportunities and discussion. First, I mainly focus on the complete ex-post assessment of the analysed regional development policies. As a consequence, state of the art machine learning and deep learning techniques may be used in future research works to forecast the effects of alternative policy measures and design interventions minimizing the misuse of resources. Second, the majority of my analyses is conducted at EU level and considers heterogeneous sectors. More vertical analyses on specific countries or sectors may provide additional insight on the factors that may influence the impact of the EU Cohesion Policy in a specific context or sector. Properly addressing such limitations in future research studies, may allow to generate additional value and empirical evidence in support of a proper design and implementation of regional development policies.

# DATA SHARING: AN ENABLING FACTOR FOR THE DEVELOPMENT OF DATA DRIVEN SERVICES IN THE EMERGENCY CONTEXT

**Valeria Maria Urbano** - Supervisor: Giovanni Azzone

Co-Supervisor: Piercesare Secchi

The increasing volume of data generated by individuals and organizations combined with the technological development of storing and processing data systems provide significant growth and innovation potential. Data-driven innovations have the potential of generating enormous benefits by enabling several applications in different fields. For example, data data-driven services can improve healthcare through personalized medicine, create new mobility solutions and improve sustainability and efficiency.
The growing attention towards the opportunities provided by data led organizations to explore the possibility of accessing other sources of data seeking new untapped potential deriving from the combination of different data sources. The relevance of data sharing was further exacerbated in emergency context characterized by profound and rapid transformation. Having access to data may support the development of insights relevant to the understanding of customers, businesses, markets and the external environment. By increasing visibility, these insights can enhance the development of proactive response plan, hence increasing responsiveness of the organization. The importance of data sharing during emergencies had already emerged in the last decade during health emergencies, as the Ebola and the Zika virus disease outbreaks, when researchers started rapidly sharing epidemiological data in public repository. The recent COVID-19 outbreak confirmed the importance of data for increasing situational awareness, hence enabling effective and timely decision-making. By sharing information, from the initial genome sequencing of the coronavirus to the incubation period and effectiveness of safety measures, the international data sharing practices put in place among researchers significantly contributed to scientific progress. The COVID-19 emergency also revealed the importance of exaptation of technology for increasing responsiveness to the challenges, hence supporting crisis management. However, despite the undergrounded optimisms around data sharing practices, these initiatives often fail. Data sharing can result in risks including loss of control on the data, protection of privacy and security. One of the major ethical concern is given by the fact that secondary data users might not respect the confidentiality measures agreed with data owners. The risk and the ethical concern is further exacerbated when individual level data are shared, hence triggering the risk of identification of individuals. A further source of concern arises when considering the whole data lifecycle including the usage of data shared by other parties for decision making. During the whole process that led to the transformation of data into actionable knowledge, the way data and information are governed may lead to the development of low-quality information products or services. The need of promptly implementing data sharing initiatives during COVID-19 also posed new challenges. Recent research studies focusing on the topic of data sharing during the outbreak provided first attempts of analysis of barriers and enablers. Authors pointed out that a major issue during the implementation of such initiatives in short periods is related to homogenization of data standards and structures. The health emergency further exacerbated the lack of resources for collecting, structuring and managing data throughout the different stage of the data sharing processes. The need of implementing data sharing processes in short period of times posed also new challenges to the level of privacy and security of the data sharing and management activities. Data sharing during the health emergency required the implementation of procedures that guarantee security and privacy while allowing prompt data sharing.

A better understanding of the data sharing process and the dynamics characterizing the initiatives implemented in the real world can help to find mechanisms that strike a balance between risk and rewards by increasing the value of the data unveiling the opportunities provided by the collection, integration and usage of data, while also mitigating potential risks. Addressing this need, the aim of the study is to analyse data sharing processes in the context of the emergency with a view to identifying opportunities provided by data sharing and major challenges to be faced. The thesis comprehends four chapters dedicated to the analysis of data sharing, considering the extended definition of the concept. Considering the entire life cycle of data flow, data sharing includes three phases: deposition, integration and translation. It is worth to highlight that the value of data does not lie only on the sharing activity but rather on the usage of data by other organizational groups. Unveiling the opportunities provided by the data sharing process requires organization to consider the whole data value chain when implementing data sharing initiatives. In general, data sharing can be broken down into a three-phase process including a data deposition, integration and translation phase. The first element refers to the provision of data made accessible to other organizations. Secondly, the integration involves the combination of data coming from different sources into a database that provides a unified view of them. Thirdly, the last element, i.e., data translation, refers to the last phase of the data lifecycle translating data into effective use by multiple stakeholders. The data sharing process is not, indeed, an objective but rather a mean to translate data into actionable information.

The research study comprehends four independent but interrelated chapters focusing on the different stages of the data sharing process and framed into two research settings: the healthcare and the mobility sectors. The choice of these two sectors is driven by the level of complexity characterizing data sharing initiatives and by the relevance of data sharing processes respectively. Regarding the former, the level of complexity is given by the number of actors involved and the type of data that is object of sharing, which is often time, location and identity dependent. The structure of the thesis reflects the need of analyzing data sharing processes in different organization and data settings. The four chapters are dedicated to different organizational and data settings to gather insights on specific dynamics that characterize diverse contexts. From an organizational perspective, three different settings are defined according to the actors involved in the process: i) intra-organizational, ii) inter-organizational, and iii) social data sharing. From the data perspective, two different data settings were analysed: i) time and position-dependent data and ii) time, position and identity-dependent data.

It is worth mentioning that specific research questions addressed in each chapter were motivated by real-world problems that emerged during the COVID-19 pandemic. Aside from the theoretical contributions of the thesis, the definition of research questions stemming from real-world problems contributed to the achievement of findings with proven practical implications.