MACHINE LEARNING APPLICATIONS IN ECONOMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Qilu Yu

May 2022

© 2022 Qilu Yu

MACHINE LEARNING APPLICATIONS IN ECONOMICS

Qilu Yu Ph.D.

Cornell University 2022

There are three chapters in this dissertation. Chapter 1 introduces the machine learning and its advantages and disadvantages in the context of economic research. The machine learning algorithms can complement traditional econometric methods and expand the boundaries of research.

Chapter 2 uses a novel deep learning approach, “Temporal Causal Discovery Framework (TCDF)” to uncover the causal graph structure on the European countries’ credit default swaps during the 2010-2013 eurozone crisis. TCDF uses attention-based convolutional neural networks combined with a causal validation step to learn the causal relationships and the time delay between a cause and the occurrence of its effect. This study provides a granular report of the eurozone crisis contagion and spillovers, adding new findings to the repository. The benchmark Granger causality tests are implemented by vector autoregression. The comparison between the two methods suggests the TCDF can filter the “real” cause-effect relationships using causality validation.

Chapter 3 extends the famous Jordà-Schularick-Taylor Macrohistory database with a new crisis variable by referencing other crisis datasets. This new dataset contains 1570 observations of 17 countries from 1870 to 2016, of which 322 observations are crisis periods. XGBoost, random forest, and the logit model are applied to this dataset to establish early warning systems for financial crisis. Though XGBoost is a popular tool in applied ML, it has rarely been used in previous studies for early warning

systems. This chapter shows that XGBoost outperforms the benchmark logit model, its performance is on a par with random forest. The two machine learning methods can achieve excellent prediction performance evaluated by the AUROC. Shapley values of the variables are calculated from the models to rank the variable importance in terms of predictive power.

BIOGRAPHICAL SKETCH

Qilu Yu was born in Huangmei, China in 1987. She moved to Wuhan at the age of 9 and finished her pre-college education in Wuhan. She entered Wuhan University in 2005 and earned a dual degree of Bachelor of Arts in Economics and Bachelor of Science in Mathematics in 2009. Later, she started her graduate studies in the Department of Economics at Cornell University. She holds a Level 3 Award of Wines from WSET, and is also an enthusiastic graphic designer.

This document is dedicated to my parents, Manli Qi and Jian Yu.

ACKNOWLEDGMENTS

I am truly grateful for my committee members who have helped and supported me throughout my long journey in pursuing my degree. Being an “outlier” student, I must have brought so much troubles, but they never gave up on me, yet always encouraged me during times of difficulties. My advisor, Prof. Yongmiao Hong, has helped me so much in every way, without Prof. Hong’s guidance and support, I would never have the chance to write this thesis. With his vast knowledge, hardworking spirit and altruistic kindness, Prof. Hong has always been the role model I look up to and learn from.

I am also deeply indebted to Prof. Nancy Chau, who had supported me unselfishly during my hardest moments. Prof. Chau’s kind words and actions had given me faith and confidence when I was low and in distress. Prof. Panle Barwick’s genuine support has motivated me enormously. With her guidance and encouragement, I was able to finish writing the thesis.

I also want to express my gratitude to Prof. Assaf Razin and Prof. Karl Shell, who had taught me not only economics but also perseverance and determination. I want to thank Ms. Robin Hamlisch at Cornell Health, who had brought hope, and given me the whole-hearted understanding.

My special thanks go to Prof. Hong’s wife, Xin Wang. Her kindness and cheering character are like the warm sunshine in Ithaca’s winter, that always inspires me. I also want to thank Prof. Jimmy Yu, Dr. Ming-Yi Chou, Cynthia Du, Prof. Paul Koch and many others who helped me along the way.

Last, I want to thank my parents for all their love, patience and sacrifice.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Leo Breiman, one of the founding fathers of modern machine learning, writes in his farsighted 2001 paper: “There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that a given stochastic data model generates the data. The other uses algorithmic models and treats the data mechanism as unknown” (Breiman, 2001). The ideas conveyed in Breiman’s work show the differences in model-driven approach and data-driven approach.

Machine learning algorithms are flexible data-driven models, they differ from the traditional econometric and statistical tools in many aspects. In this chapter, we provide a brief overview of machine learning and explain why it is suitable for some economic questions. In Chapter 2 and Chapter 3, we apply machine learning methods to two economic questions.

1.1 Development of machine learning

This section will briefly introduce the concepts and terminology of machine learning (ML), its early development and connection with the econometric toolbox, then we lay out the foundations for the burgeoning development of ML in recent decades. The goal of this part is not to conduct a comprehensive investigation on ML, but to provide the background and intuition of applying ML in economic research, especially on the empirical side.

The first task is to understand what is ML? ML is a term that often comes with Artificial Intelligence (AI) and Deep Learning (DL). In fact, these three terms can be put in a hierarchical order, that DL is part of ML, and ML belongs to the broader concept of AI. AI refers to systems or machines that mimic the problem-solving and decision-making capabilities of the human intelligence and can iteratively improve their abilities based on the data (Russell, and Norvig, 2021). Narrowing down from the general ability to emulate the human mind, ML, which is a subcategory of AI, primarily deals with pattern identification and decision making through experience and data.

ML is a broad topic that encompasses computer science, statistics, engineering and other fields; hence, its definition is context specific. From the perspectives of an applied economist, Athey (2018) provides a relatively "narrow and practical" definition of ML: "machine learning is a field that develops algorithms designed to be applied to datasets, with the main areas of focus being prediction (regression), classification, and clustering or grouping tasks." Athey confines this definition to two major branches in ML that are most commonly used in economics and other social sciences: supervised learning and unsupervised learning. DL is a subset of ML, which does not fall directly under the two branches of supervised learning and unsupervised learning. DL can be used with or without supervision. DL enables the computer to learn complicated concepts using a hierarchy of concepts, with each concept defined or learned through previous simpler concepts (Goodfellow et al., 2016). Recent advancements in ML are mostly in the DL area (LeCun et al., 2015).

ML tools can be divided into three subcategories: supervised learning, unsupervised learning, and reinforcement learning. The differences of the subcategories lie in the goal of optimization and data type. In supervised learning and unsupervised learning, the goal is to minimize a loss function associated with prediction performance, whereas, in reinforcement learning, the goal is to maximize the reward in a dynamic environment. Between supervised learning and

unsupervised learning, the key difference is the data type. Supervised learning uses labeled datasets to train or "supervise" the algorithm. The machine can measure its prediction accuracy using the labeled input-output pairs, thus learn from its experience. Unsupervised learning analyzes and groups unlabeled data by detecting similarities and patterns in the data. The machine needs to discover the hidden structure of the data by itself, hence "unsupervised".

Supervised learning and unsupervised learning are the major tools that are employed in recent economic studies, and are also the focus of this paper. Supervised machine learning tries to identify a function that maps a known input to an unknown output based on example input-output pairs (Russell and Norvig, 2021). A supervised learning algorithm infers a function by analyzing a training data set containing labeled examples, the function can produce an output from unseen data (Mohri et al., 2012). The nature of supervised learning is optimization with regularization using computer algorithms (Hong and Wang, 2021). Supervised learning mainly deals with two types of problems, regression and classification. In regression problems, the output is a continuous variable (for instance, housing prices), common models include linear regression and polynomial regression with shrinkage methods. In classification problems, the output is a categorical variable (for instance, crisis or no crisis), common models include support vector machine, trees-based models, and neural networks.

Unsupervised learning can analyze the unlabeled, unclassified data to detect similarities and patterns, thus cluster or group the data. Unlike supervised learning, unsupervised learning methods cannot be directly applied to a regression or a classification problem, because the unlabeled data does not have input-output pairs. Common usages of unsupervised learning are clustering, anomaly detection and dimensionality reduction. Clustering techniques such as the K-means algorithm are used to partition the feature space into subspaces, then the clusters can be used to create new features based on subspace membership (Athey, 2019). Anomaly detection is to examine the data to discover atypical data points (e.g., fraud detection in

insurance claims). Dimensionality reduction is an important technique in preprocessing high-dimensional data (e.g., images, texts, video footage, etc.) with a large number of features. It can reduce the number of features to a manageable size without hurting data integrity. For example, principal component analysis (PCA) is a common technique for dimensionality reduction. Unsupervised learning techniques are often used in different stages of complex ML problems, such as reducing data dimension in the preprocessing stage. Some ML algorithms combine both supervised learning and unsupervised learning, for example, the Generative adversarial networks (GANs, Goodfellow et al., 2014) contain two sub-models, a generator model for generating new examples and a discriminator model to classify whether generated examples are real or fake, the discriminator model is unsupervised learning. In such cases, the learning method is called semi-supervised learning.

Reinforcement learning is the training of a computer agent to make a sequence of decisions through repeated trial-and-error interactions in a dynamic reward system. The machine is trained without answers or hints. It can only learn from its experience, and its goal is to maximize the total reward. Examples of reinforcement learning include Alpha Go and self-driving cars. This "game simulation" feature of reinforcement learning provides functionality in optimal control problems in economics, game theory, operational research and finance (Charpentier et al., 2020), but it is of less importance to causal inference and prediction problems, which are the emphasis of most economic studies.

The above part gives a concise introduction to the concepts and terminology of ML, next we will discuss the early development of ML and its connection with the econometric toolbox. Although ML seems like a new concept, it is not a new trick for researchers. ML has been around since the 1950s. Pioneering ML research was conducted since 1950 (Russell and Norvig, 2021). In 1950, Alan Turing was the first to propose the question "can machines think?" In 1951, Marvin Minsky and Dean Edmonds created the first neural network. In 1959, Arthur Samuel

popularized the phrase “machine learning”. Since then, researchers have developed more and more ML algorithms. For example, the nearest neighbor was invented in 1967, recurrent neural network was created in 1982, reinforcement learning was discovered in 1989. In fact, a lot of the well-known ML algorithms are decades old.

Econometricians, such as Halbert White, are also pioneers in the development of ML algorithms. White applied neural networks in economic research to predict the IBM stock market prices (White, 1988), and had done extensive research on neural networks. Traditionally, econometricians usually adopt statistical tools to build probabilistic models to describe economic phenomena (Charpentier et al., 2019). The model free ML algorithms are unsuited for this purpose due to its lack of interpretability, therefore ML are rarely used in economic research. In econometrics, there is a line of research that exercises similar methodology to ML, that is nonparametric analysis. Nonparametric analysis uses particular statistical techniques that do not require pre-specified functional forms for objects being estimated, nor require much prior information on the data generating process (DGP, Racine, 2008; Cai and Hong, 2003). If in a regression framework, it is called “nonparametric smoothing”. Nonparametric smoothing includes spline smoothing, kernel smoothing, K-nearest neighbor (KNN) smoothing, and decision trees, to name but a few.

Nonparametric analysis is analogous in spirit to ML algorithms, that they both assume the DGP is an unknown stochastic process. They both depend on the data to derive the forms of the estimator or predictor, thus provides little economic interpretability. Nonparametric analysis generally requires a large dataset for precise estimation because there are many unknown parameters, it can suffer from the curse of dimensionality when using local smoothing methods for multivariate data (Hong, 2020). Some ML algorithms (e.g., K-nearest neighbors (KNN), decision trees, neural networks) can be viewed as nonparametric methods with a regularization component based on optimization using computer algorithms. The regularization can afford

more explanatory variables, thus "break" the curse of dimensionality (Hong and Wang, 2021). Sarle (1994) has shown that the multilayer perceptrons, which are the most commonly used neural networks, are just nonparametric nonlinear statistical models. In this regard, econometricians are no stranger to the ML concepts and methodology.

Just like nonparametric analysis, ML algorithms require large datasets and strong computational power to fulfill their full potential. Up until the 1990s, datasets were small, computers were slow and costly, those impediments hindered the development of ML. Though ML algorithms have been around since the 1960s, their broad applications only start in the early 1990s. This late resurgence is brought by three forces: big data, cheap computational power and algorithmic advances.

First, data availability increases unprecedentedly since the early 1990s. With the Internet of Things (IoT), data is created in the forms that are never seen before, such as remote sensing data, high frequency trading data, online sales data, the data generating speed also increased rapidly. In economics, areas such as agricultural economics, environmental economics, and marketing have benefited from abundant data sources to extract information and patterns in all sorts of activities. However, as the traditional statistical tools are mostly developed for smaller datasets, they lack the ability to handle the complex large datasets. The advance in big data demands new methods for data processing and analysis, the ML algorithms can precisely fill this gap. Big data and ML complement each other that ML needs large datasets to train the machines, big data can use ML algorithms for in-depth data mining. We will reserve the more detailed discussion on big data in the next section.

Second, computational power has experienced an exponential growth over the past decades[1], innovations such as parallel computing increase the speed and efficiency of central processing

[1] According to the Moore's Law, which states that the number of transistors on a microchip doubles about every two years.

unit (CPU) usage, and lower the cost of computing. Since the early 2000s, the computing power of multi-processor graphic cards (or GPU) is also employed by the ML community (Storm et al., 2020). All these advancements contribute to the rapid growth of ML. Third, the research community of ML from both academia and industry is constantly advancing the frontiers of ML algorithms. The open-source nature of the ML programs (e.g., R, Python, etc.), off-the-shelf ML algorithms and libraries, encourages the broad applications of ML methods (Schmidhuber, 2015). These three forces jointly create a huge pool of ML tools readily available for researchers in all fields of study.

The above part summarizes the early development of ML, the next section will discuss the strengths of ML and its comparison with the econometric methods.

1.2. Why use machine learning in economics?

ML methods are flexible, rich, data-driven models. This section intends to lay out the five reasons why ML methods are gradually recognized in economic research. For comparison, we will use traditional econometric methods for benchmarking. ML methods and econometric methods have different objectives. ML methods are primarily intended for accuracy in prediction and classification, while the econometric methods are usually developed for deriving the statistical properties of estimators. One important terminology distinction arises from this difference in objectives. In ML, terms such as bias, standard deviation or mean squared error are defined for the prediction process, while in econometrics, those terms are reserved for the coefficient estimators in hypothesis testing, the statistical properties of estimators are usually not obtained in ML (Storm et al., 2020).

Despite the differences, one should also acknowledge the similarities between ML methods and econometric methods. They are not competitors, but collaborators. They excel in solving

different types of problems, therefore can complement each other to produce high-quality research. The ideas in this sector are drawn from previous works such as Athey and Imbens (2017), Athey (2018), Hong (2021), Hong and Wang (2021) and many others. The reasons provided here are not a comprehensive overview of ML's merits, but they build the foundation for discussions in chapter 2 and 3 of this paper.

1.2.1 Big data promotes the use of machine learning

The first reason is the impact of big data. Traditionally, economists are accustomed to work with data that fits nicely in a spreadsheet, but the big data are often too large and complex for a spreadsheet. The quality and quantity of economic data are expanding rapidly (Einav and Levin, 2014). The use of big data can improve causal inference and provide better prediction of economic phenomena (Harding and Hersh, 2018).

Big data is commonly characterized by the 4 V's: volume, velocity, variety and veracity (Hong and Wang, 2021). Volume comes from the word "big", which shows the sheer volume. Coming into the information age, humans are creating as much information as once did from the dawn of civilization up until 2003 in every two days (Eric Schmidt, the CEO of Google). We now have data that is several gigabytes in size. In the past, data was usually collected for a specific purpose by a national statistical agency; but as the world becomes increasingly quantified, data are now collected through a vast ecosystem of software and hardware, including phones, Wi-Fi connected appliances, and satellites (Harding and Hersh, 2018). Through collaborations, economists can study many large-scale proprietary and administrative data, for instance, the eBay online audition data and tax records.

Velocity means high frequency. In economics, high frequency data (i.e., intraday trading data) has been widely used for prediction problems in the past decades, but the usage is mostly

in finance. In other fields such as macroeconomics, the data is usually aggregated and in low frequency, new initiatives are creating more high frequency data in these fields. For example, FRED-MD[2] is a large macroeconomic monthly database that includes 127 granular variables such as retail and services sales, it is updated in real time for economic nowcasting. Researchers at MIT and Harvard put together the Billion Prices Project[3] to offer high-frequency online retail price, as a proxy for the Consumer Price Index, to "bring big data for macro and international economics".

Variety stands for the numerous types of data. As more information can be digitized, data such as text, audio and video, satellite images, can all be captured and stored in relational databases. Unlike traditional data, those non-traditional data is often unstructured. Lots of studies have used unstructured data in areas such as health economics, agricultural economics, environmental economics. For example, Kleinberg et al. (2015) use ML methods on a dataset that contains 3,305 explanatory variables to study mortality rate of patients after surgery. The 3,305 variables include image data (MRI scans), demographic data (age, geography, etc.), interval data (range of blood pressures), etc.

Veracity points at the noise, abnormalities, and inconsistencies in the data. There are two levels of veracity in big data. One is data credibility, meaning the reliability of the data sources and collecting process. The other is the discrepancies in large datasets where the data are sourced differently and are in different structures. Traditional econometric methods have low tolerance towards data veracity, while ML methods can cope well. For example, some ML algorithms, such as the K-nearest neighbors and naïve bayes, are robust to missing values in the dataset.

[2] See https://research.stlouisfed.org/econ/mccracken/fred-databases/. FRED-MD and FRED-QD are large macroeconomic databases designed for the empirical analysis of "big data."
[3] See http://www.thebillionpricesproject.com/

There are also ML based imputation methods which can impute missing values without the assumption of normality or specification of a parametric model (Hong and Lynn, 2020).

With the 4 V's, big data has expanded the boundaries of research in many fields of study, such as geology, hydrology, and biology. The enriched data sources enable in-depth analysis and can extend the research topics. But big data is not just expanded larger datasets, it has a number of properties that differ from traditional data, which requires different practice of data storage, management, processing and analyzing. Here, we pick out three properties that are most relevant to ML algorithms.

First, big data often comes unstructured. Unlike the traditional data types (time-series, cross section, and panel data) which fit well in spreadsheets, big data has novel date types such as speech, text, images. Those unstructured data types cannot be directly processed using traditional statistical tools to extract useful information. In practice, researchers need to turn those unstructured data into quantifiable indices or metrics based on domain knowledge. However, this data transformation process may cause bias and loss of information. In contrast, ML methods can process unstructured data directly, or use dimensionality reduction techniques to obtain data in lower dimension while preserving data integrity. This automatic data-driven feature extraction is crucial to the avoidance of introduced bias and keeping as much information as possible. For instance, principal component analysis (PCA) is a widely used dimensionality reduction technique. PCA is a ML technique that transforms a large number of features into a smaller set of uncorrelated features called principal components. For complex unstructured data, ML methods have been proven to be of great importance for data preprocessing.

Second, big data can have a large number of explanatory variables (K variables). For example, in marketing, if location data such as zip code is to be used, thousands of dummy variables would be included; in environmental economics, climatological data can store numerous features for one observation. If the number of observations is N, when K is large or

$K > N$, the inclusion of a large number of explanatory variables would lead to the curse of dimensionality for traditional econometric methods. For instance, regression or classification become susceptible to error and overfitting as the data space becomes sparse with a large number of explanatory variables. To use big data that has a lot of explanatory variables, the traditional econometric approach focuses on a subset of variables by certain criteria, or uses aggregation based on domain knowledge. Similar to the unstructured data, this process may introduce bias and cause loss of information. ML methods are much more flexible in variable selection and dimensionality reduction through data-driven feature selection. PCA is again a popular tool for this task. Spike-and-slab method in the Bayesian structural time series (bsts) model is another technique to select features that have higher predictive power. Random forest (and other tree-based models) can overcome the curse of dimensionality by building independent decision trees that are only trained on a subset of features. ML methods are more flexible and efficient when dealing with a large number of explanatory variables.

Third, big data usually have complex data sources with large heterogeneity, missing data points and noises. Nonlinear relationship and multicollinearity among the explanatory variables are also common. Those issues would render most econometric methods inapplicable, methods such as generalized additive models, nonparametric methods (kernel estimation) can be put to work, but they are subject to a manageable size of explanatory variables. In contrast, ML methods show great tolerance for those data issues. Many ML methods can identify nonlinear and complex linear connections in the data, for example, K-nearest neighbors (KNN) uses the average value of k nearest neighbors for the predicted outcome; decision trees follow the tree-like structure to split data at each node; these algorithms are well suited for nonlinear datasets. Multicollinearity is a challenge in traditional multivariate regression and should be avoided for the sake of interpretability. But in ML, when the goal is prediction, multicollinearity does not inhibit ML algorithms' ability to achieve prediction accuracy. Tree-based models such as

random forest and XGBoost are immune to multicollinearity, because the algorithms only pick one of the correlated features for deciding a split at a node. When dealing with noisy data, many noise identification and noise handling techniques in unsupervised learning can be used to filter out the anomalies and improve prediction accuracy (Gupta and Gupta, 2019). Also, many ML algorithms (KNN, naïve bayes, etc.) are robust to missing data points.

In a nutshell, with the advent of big data usage in economic research, ML methods quickly gains attention. The flexible ML algorithms are designed for large datasets and can dig into rich data to decipher complex data patterns. Not only can ML methods be used for prediction and classification problems, they can also complement the traditional econometric methods through techniques such as dimensionality reduction and variable selection.

This part summarizes the impact of big data on ML. The next section will explain the differences between the two research paradigms and show the merits of data-driven ML methods.

1.2.2 The two research paradigms

At the beginning of this chapter, we introduced the two cultures in using statistical modeling to reach conclusions from data (Breiman, 2001), one is model-driven, the other is data-driven. The two cultures are exactly the two scientific research paradigms described in E (2021): the Newtonian paradigm and the Keplerian paradigm. These two paradigms encompass almost all modern scientific research.

The Newtonian paradigm, or model-driven approach, discovers the fundamental principles that govern the world and the universe. Researchers distill the relevant factors from countless causes and conditions, build theories and models to explain the observed phenomena, then use data to validate the framework and estimate model parameters. The Keplerian paradigm, or data-

driven approach, observes the data to induce and generalize scientific discoveries. Researchers make no assumption of the underlying data generating process, they analyze and approximate the data to develop practical understanding of the world and the universe.

In economics, the Newtonian paradigm prevails since Adam Smith explained the theory of the invisible hand in his "Wealth of nations". Economics is a discipline that aims to study the underlying relationships among economic factors, to discover how the factors interact with each other to bring about an economic phenomenon. The modern economic research in the Newtonian paradigm can be characterized using the following process (see Hong (2007) for a detailed exposition). The first step is to collect data and summarize the empirical stylized facts. For instance, the positive relationship between years of education and lifetime income is a stylized fact. The second step is to build an economic theory or model to fit the stylized fact. Economists usually use mathematical tools to set up models that show how the variables work together to produce the observed outcome. The third step is to test the model using econometric tools, to validate model specification and estimate parameters from data. Finally, the validated model can be used for decision-making, economic forecasting, etc. To sum up, model-driven economic research aims to uncover the economic principles by building theories and models, then use econometric tools to validate the models and quantify the relationships among economic factors from data.

An early example of the model-driven approach is the Solow growth model built by Nobel laureate Robert Solow. It uses a Cobb–Douglas production function to show how the long-run economic growth is governed by capital accumulation, population growth, and increases in productivity. Solow (1957) applies this model to the United States' data and finds that productivity increase brought by the technological progress is the main driver for economic growth. With the advance in mathematical modeling, statistical tools and data availability, new theories and models are constructed with more variables and complex functional forms. For

instance, the dynamic stochastic general equilibrium (DSGE) models attempt to explain and predict the co-movements of the macroeconomic time series over the business cycle by applying the micro-foundations of a competitive equilibrium model. A typical DSGE model includes a representative household, a representative firm, the government, and foreign sectors. Each agent has its own set of maximization problems which interdepend on other agents in the model. Shocks are also introduced to inject randomness. The DSGE models are used to make economic predictions about the business cycles and economic growth. The complexity demonstrated by the DSGE models shows the direction of evolvement in the model-driven approach.

Though DSGE models enjoy much popularity among macroeconomists, they also received critiques. The many assumptions about a representative household with an infinite lifetime and competitive market are targets of criticism (Stiglitz, 2018). Robert Solow points out that small deviations can be amplified through the complex systems in the DSGE model, which can lead to substantive digression from the true outcomes (Solow, 2010). Some DSGE modelers respond to the critiques by increasing the level of complexity, for example, by introducing heterogeneous agents (Christiano et al., 2018). This trend of increasing complexity is common in many economic models.

The Keplerian paradigm, or data-driven approach aims to find the relationships between the explanatory variables and outcome variables of an economic phenomenon, with few or no prior assumptions. The data-driven approach usually has three steps. The first step is to collect data and summarize the empirical stylized facts (the same as the model-driven approach). The second step is to use data-driven models to fit the data. For example, one can use the autoregressive integrated moving average (ARIMA) to model the monthly totals of airline passengers, or train the XGBoost algorithm to predict daily stock market price. The third step is to use the model to make predictions on new data. The model usually lacks the ability to interpret the relationships among variables, and requires a sufficiently large dataset. Since economists are always

interested in finding causal relationships, and the data availability is not ideal in the past decades, the data-driven approach has been a less popular subject in economics.

Generally, there are two types of data-driven approaches in economics. The first type is statistical or econometric approach. For example, many time series forecasting methods, such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA) can predict future values of a time series based on past results without building structural models based on domain knowledge. Nonparametric analysis is another example of model-free statistical tool. The second type is ML methods. ML algorithms can learn from data and experience to identify patterns and relationships between variables with little prior information.

To sum up, the model-driven approach aims to identify important variables and estimate the parameters of a model that describes the distribution of a set of variables (Athey and Imbens, 2019), the goal of which is to understand the principles of economics. The data-driven approach puts the emphasis on extracting patterns and information from data to develop practical understanding of the underlying data generating process.

If we assume that the economy is governed by certain stochastic data generating processes (this statement itself, is an assumption), one can approximate the real economic activities using mathematical or statistical modeling on a set of variables, and disregard less relevant ones. The simplified models can help to understand the main aspects of the DGP; however, no model can ever provide a completely accurate depiction of the DGP, that "all models are wrong but some models are useful."[4] If the model is a reasonably accurate approximation of the DGP, then model outputs are also approximately true (Hong, 2020). If the model is mis-specified, such as leaving out important factors, including irrelevant ones or choosing inappropriate function forms; the

[4] A famous quote by statistician George Box.

model outputs can be biased and inconsistent. Model mis-specification can have adverse effects in policy analysis and decision making, this is a major drawback of the model-driven approach (Hong, 2021).

Also, we are often very willing to make strong assumptions when constructing a model, but the strong assumptions cannot be directly tested, and they might just be the source of bias. If the model assumptions are to be relaxed, like the case with DSGE models, where modelers introduce heterogeneous agents to relax the representative agent assumption, model complexity increases. If we introduce nonlinearities, interactions or heterogeneity to the model, the model will become complex and susceptible to bias, model interpretation will become more difficult as coefficients cannot be interpreted directly. We often accuse the flexible and complex ML methods for their lack of interpretability, however, this tradeoff between flexibility and interpretability is not exclusive to ML, it is also a tradeoff for the structural models (Storm et al., 2020). Simplified economic structural models cannot capture every aspect of the economy, and building complex models "would make them unwieldy for either theoretical insight or applied analysis" (Low and Meghir, 2017).

Another drawback of the model-driven approach is model uncertainty. In practice, researchers would consider many model specifications and perform various specification tests before choosing a preferred model that can produce preferable results. The standard practice is to present estimates of the preferred specification with several other specifications with different functional form, controls or instrument variable. This is to show that the estimate of the parameter of interest is not very sensitive to the choice of the preferred specification (Athey and Imbens, 2017). The fact that an estimated parameter varies with different models represents a simple form of model uncertainty (Varian, 2014). This approach of picking the "best" model out of many model specifications cannot deplete all possible models and functional forms, therefore the results can suffer from model uncertainty (Varian, 2014). In contrast, ML

algorithms can afford highly flexible functional forms and use data-driven function selection to avoid this problem.

In summary, model-driven approach and data-driven approach differ in their objectives and methodologies, each approach has its advantages and limitations. Model-driven approach is not immune to bias and errors, and is not always superior to data-driven approach. Therefore, it is important that we recognize the virtue and strength of the data-driven approach (especially the ML methods).

1.2.3 Integrative modeling

The last section shows the dichotomy of the model-driven approach and the data-driven approach, but there are other classifications as well. Hofman et al. (2021) point out that empirical modeling can be organized along two dimensions, representing different levels of emphasis placed on explanation and prediction (Table 1.1).

Descriptive modeling (quadrant 1) refers to activities that define, measure, collect, and describe relationships between variables of interest. Explanatory modeling (quadrant 2) refers to activities to identify and estimate causal effects, not focusing directly on prediction. A lot of model-driven economic research belongs to this quadrant. Predictive modeling (quadrant 3) refers to activities in predicting the outcome of interest directly but do not explicitly deal with the identification of causal effects. Most ML methods belong to this quadrant. Integrative modeling (quadrant 4) combines the explanatory properties of quadrant 2 and the predictive properties of quadrant 3, it refers to activities that attempt to predict unseen outcomes in terms of causal relationships, the focus is on generalizing “out of distribution” outcome that might change naturally, or through human intervention such as controlled experiment (Hofman et al., 2021).

Table 1.1 A schematic for organizing empirical modeling along two dimensions, from Hofman et al. (2021)

	No intervention or distributional changes	**Under interventions or distributional changes**
Focus on specific features or effects	*Quadrant 1:* Descriptive modeling Describe situations in the past or present (but neither causal nor predictive)	*Quadrant 2:* Explanatory modeling Estimate effects of changing a situation (but many effects are small)
Focus on predicting outcomes	*Quadrant 3:* Predictive modeling Forecast outcomes for similar situations in the future (but can break under changes)	*Quadrant 4:* Integrative modeling Predict outcomes and estimate effects in as yet unseen situations

The concept of integrative modeling admits to the tradeoff between explanatory insight and prediction accuracy, but it also recognizes this tradeoff as "an exciting opportunity for new and impactful research" (Hofman et al., 2021). For instance, in estimating average treatment effects, economists have been using semi-parametric methods without making parametric assumptions about how explanatory variables affect the outcomes in the 1990s (Athey, 2018). Since the 2010s, economists introduce ML methods into this framework. Belloni et al. (2014) uses a double-selection method based on LASSO in high-dimensional data with many explanatory variables, to allow inference about economically interesting model parameters. Athey et al. (2018) propose to use "residual balancing" that takes the average of the efficient score as the measure of the treatment effect, which is calculated from the inverse of estimated propensity score and the conditional mean of the outcome variable. The propensity score weights are obtained through quadratic programming and the conditional means are estimated using LASSO. Their main result shows that this procedure is efficient and can achieve the same rate of convergence as a well specified structural model.

The above examples are situations when we are interested in estimating a parameter of causal interest, but the tools we use to recover that parameter may contain a prediction component (see

Mullainathan and Spiess (2017) for a review). Another approach in integrative modeling is akin to a "coordinate ascent" algorithm, where researchers iteratively alternate between predictive and explanatory models during the experiment stage and data analysis stage (Hofman et al., 2021). In experiments, researchers can use ML algorithms to predict participants' decisions and identify features of importance, the results can then be used in new rounds of experiments to test the predictions, and so on.

In conclusion, explanatory modeling is usually used to identify causal relationships among variable, such as quantifying policy impact and estimating counterfactual outcomes, predictive modeling (mainly the ML methods) can complement the explanatory models through innovative methodology for predictive accuracy, combining those two modeling strategies can expand the horizons of economic research.

1.2.4 Some economic problems are predictive

Conventional econometric methods are not designed for prediction problems, these model-driven methods put their emphasis on statistical inference on parameters of causal interest, rather than predictive performance. Though a lot of the economic problems that interest the researchers and other stakeholders are causal, but not all important economic problems are causal, some are predictive in nature. For predictive problems, ML methods are perhaps second to none. The primary focus of supervised learning is on accurate prediction, especially the out-of-sample predictive fit. Unsupervised learning can help with the predictive task through data-driven feature selection and dimensionality reduction techniques. Shrinkage methods, tree-based models, and neural networks are the prevailing ML tools researchers use for predictive regression and classification problems.

Kleinberg et al. (2015) argue that many important policy problems are essentially prediction problems, that "causal inference is not central, or even necessary." They study the impact of a resource allocation problem in joint replacement surgery for population with Medicare in the U.S., the benefit of surgery is an improved quality of life, but the cost is hard to quantify. The post-surgical recovery can be painful and takes several months afterwards. A key challenge is the possible mortality for individual patient after surgery. The policy decision of this study is whether the surgery on the predictably riskiest patients will be futile. The authors use 65,395 observations to fit the models using LASSO regression, then measure the payoff of the surgeries on the remaining 32,695 observations. Results show that, for each year, in the whole Medicare population, replacing the riskiest 10 percent patients with lower-risk patients according to ML predictions would avert 10,512 futile surgeries and reallocate 158 million dollars. This replacement can avoid operating on 38,533 riskiest patients who would have died within one year of surgery. This study shows how improved prediction using ML can have significant policy impacts, without identifying the causal factors. There are many resource allocation problems like this in policy problems (see Athey (2017) for a review). For instance, how to predict the locations that might need food-safety inspections? And how to send firefighters more efficiently?

ML methods can also help with pure predictive problems. Early warning system (EWS) is a predictive tool for financial crises prediction or other risk detection, there is an increasing literature on using ML methods to construct EWS for financial crisis. Real-time nowcasting of key economic indicators (mostly in macroeconomics) is another example where ML algorithm can make a difference.

In such predictive problems, the dataset under study is often large and contains a lot of unstructured variables, flexible ML methods can help researchers to better generate policy impacts and economic insights than traditional econometric methods.

1.2.5 On smaller datasets

In the beginning of this paper, we quote Leo Breiman that algorithmic models can be used not only on large datasets but also on small datasets, as a more accurate and informative alternative (Breiman, 2001). In the above sections, we have established the foundations for why ML methods are ideal for big datasets, and why ML methods perform well in larger datasets. This section will introduce some ML techniques that can cope with a limited sample size.

In economics, the majority of data in areas such as macroeconomic, is of low frequency (e.g., annual GDP) and a limited size; decades-old historical data remains scarce; and there are special datasets that only have few observations. From the frequency aspect, most weekly or daily data are sufficiently large for ML algorithms, yearly or monthly data can be a stretch. We often tend to think that ML methods are only reserved for big datasets, but researchers (mainly outside of economics) have shown that ML methods can be used on smaller datasets as well. Data augmentation techniques can be used to increase the volume and there are methodologies for special treatment.

Bootstrapping is a common statistical technique for data augmentation. In ML, a related technique is bagging, which is essentially bootstrapping aggregation. In bagging, several machines are trained on random subsets of the original dataset, then the results of the machines are aggregated to get the final prediction. The random subsets are drawn with replacement (just like bootstrapping). Bagging is a common practice for tree-based models (it is a built-in technique for random forest), it can improve the prediction accuracy by introducing randomness into the structure, then make an ensemble out of the predictions. Over-sampling is another ML technique for small datasets or imbalanced datasets, it can create synthetic data from the original data. For instance, synthetic minority over-sampling technique (SMOTE) is commonly used to increase data size, especially for minority categories in imbalanced datasets; generative

adversarial networks (GANs) can learn patterns from the original datasets and create new data that resemble the original data.

Other than data augmentation techniques, ML algorithms are adapting to smaller datasets as well. For instance, in neural networks, Olson et al. (2018) decompose the final layer of the fitted neural networks into ensembles of low-bias sub-networks, those sub-networks are relatively uncorrelated, therefore enabling the implicit regularization mechanism to avoid overfitting (a similar strategy as random forest which is an ensemble of low-bias, uncorrelated trees). They apply this approach to small datasets and show that deep neural networks can achieve superior prediction accuracy with minimal tuning. There are many applications of such ML algorithms outside of economics. For instance, Zhang and Ling (2018) use the degree of freedom (DoF) of the model to mitigate the issue of small data size in material science, Caiafa et al. (2021) review a collection of fifteen papers that use novel ML methods on noisy, incomplete or small datasets on different science subjects.

In economics, most studies that use ML methods analyze large datasets, there are few ML applications on smaller datasets. We argue that data size should not be an obstacle to apply ML methods. In chapter 3, we will use a relatively small dataset to establish an early warning system for financial crisis.

1.3 Limitations of machine learning

In section 1.2, we have shown the five reasons why ML methods can contribute to economic research. In this sector, we will introduce three limitations of ML, as well as recent developments to overcome them. ML excels at predictive tasks but it is not designed to identify the reasons, the most commonly known drawback is its lack of interpretability. Also, ML

models cannot provide coefficient estimates for causal inference. Overfitting is another challenge.

1.3.1 Lack of causal interpretation

The lack of interpretation stems from the "black-box" nature of ML algorithms, the relationships learned by the ML algorithms are complex and unreadable, this makes the ML methods untrustworthy for causal questions. For example, ML can be applied in assessing credit card applications, but it cannot provide sound reasons for its rejections. The basic framework of ML is to use historical data to discover patterns using computer algorithms, then use the learned knowledge to predict the outcome for unseen data. Patterns can show correlation, but correlation does not imply causation.

On the flip side, the fundamental problem of causal inference is that the real outcome and the counterfactual outcome cannot be observed at the same time. Good identification strategy is required for causal analysis. Randomized controlled experiments are ideal, but they are costly and take time. In quasi-experimental settings, methods like difference in difference and regression discontinuity have their own requirement for identification. Under the experimental settings, the causal models can use past data and control variables to "predict" the counterfactual outcome of the treatment group. In this spirit, ML methods can be employed for the prediction part to get the estimated treatment effect. Several off-the-shell ML algorithms are built for this purpose. For example, researchers at Google develop the CausalImpact algorithm based on Bayesian structural time series (bsts) predictions to study the treatment effect of an intervention (Brodersen et al., 2015), the Facebook Prophet is another tool for time series forecasting and treatment effect estimation.

To overcome the lack of interpretability, recent development in the intersection of economics and ML has infused causal inference into developing new ML algorithms. For example, Wager and Athey (2018) develop a causal random forest to estimate heterogeneous treatment effects which extend the original random forest algorithm to allow statistical inference. Farrell et al. (2021) construct novel nonasymptotic high probability bounds for deep feedforward neural networks which allow semi-parametric estimation and statistical inference. Several survey papers have documented this recent development, such as Ghoddusi et al. (2019), Moraffah et al. (2020), and Storm et al. (2020).

Outside of economics, interdisciplinary work between ML and causal inference has developed a subfield called causal structure learning or causal discovery. For example, Nauta et al. (2019) use an attention-based convolutional neural networks called Temporal causal discovery framework (TCDF) to detect the causal relationships in observational time series data. They construct temporal causal graphs which can determine the time lag between a cause and the occurrence of its effect. We will use TCDF in chapter 2 to uncover the cause-effect relationships during the eurozone sovereign crises from 2010 to 2013. In the temporal causal discovery literature, many tools are being developed and used in applied works (see Glymour et al. (2019) for review). We will discuss this topic in more details in chapter 2.

Though most ML algorithms aim at prediction accuracy and efficiency, we can incorporate more causal elements into the current ML framework. The goal is not to understand the "black-box" procedures, but to use novel approaches to obtain causal structure from the framework.

1.3.2 Inference on coefficients

Another criticism points to ML methods' inability to conduct inference on estimated coefficients. In nonlinear models (e.g., random forest, neural networks) there is no coefficient

that can be estimated. In linear ML models (e.g., LASSO and ridge regression), the estimated coefficients do not have the same statistical property as in traditional methods. For example, Mullainathan and Spiess (2017) report that LASSO regression can produce familiar coefficient output like traditional linear regressions, but the coefficients are biased toward zero. LASSO cannot form standard errors and confidence intervals about those coefficients neither.

This is indeed a fundamental problem of ML in the light of discovering quantifiable causal relationships. Some workaround attempts have been made to establish confidence intervals with ML algorithms. Studies mentioned in previous sections have sought to modify the algorithms to construct valid confidence intervals. Wager and Athey (2018) introduce their causal random forest, which is an average of causal trees; each causal tree is trained with a different subsample and variable space (similar to the nearest neighbor matching). They build asymptotic normality results for the treatment effect in random forest, and propose a consistent estimator for the variance, thus the confidence intervals can be computed. Brodersen et al. (2015) develop the CausalImpact algorithm based on Bayesian structural time series (bsts) to construct synthetic control and calculate the treatment effect, then use a Markov Chain Monte Carlo algorithm for posterior inference to report the pointwise 95% posterior predictive intervals of the treatment effect, the time series of pointwise intervals provides further information of the temporal evolution of the intervention.

Those attempts are still far from ideal for inference on estimated coefficients like the traditional statistical models. In particular, ML lacks the capability to conduct the null-hypothesis significance testing (NHST), but this inability to conduct NHST is not necessarily a shortcoming. Though the NHST is a widely used and well-accepted tool, it is not definitively powerful because the p-values and t-tests are designed to show that a theory is not inconsistent with the data, therefore the theory can be used as an explanatory tool. NHST is good at "disproving", but not "proving" (Hofman et al., 2021). For large enough datasets, trivially small

effects can be declared statistically "significant", because the large sample size N can reduce the standard error closely to zero, which will falsely push up the t statistics and lower the p-values closely to 0%. In other words, statistical significance is not the same as economic significance for large enough data (Hong and Wang, 2021). Since ML methods are mostly applied with large datasets, it might not be as important to obtain the statistical properties of the estimates, and we should always try to develop new methodology to quantify and test causal relationships in ML.

Current ML methods can provide a measure to evaluate the importance of explanatory variables (features) for the trained models. This is called feature importance or variable importance. ML algorithms can calculate and rank the feature importance for each explanatory variable, to show the level of importance of a certain explanatory variable in prediction. Feature importance can also help to determine if the trained model is consistent with domain knowledge.

Several approaches exist to calculate feature importance. One approach is feature permutation importance. It evaluates how the model score (or prediction error) changes when an explanatory variable is not included in the prediction. Any scoring metric can be used for measurement of the model score. If the model score shows no change when one explanatory variable is not included, it suggests that the model does not rely on this variable. In contrast, a large model score change indicates a large impact, and the sign of the change also matters (depending on the specific scoring metric). Shapley value is an extension of the feature permutation importance based on the game-theory attribution method; it takes the average of all the model score changes of a variable to all possible combinations (coalitions) with other explanatory variables.

Another approach is the impurity-based feature importance, a technique mainly used in tree-based models. Impurity is measured by the splitting criterion using a loss function at each node of the trees. The loss function can be of different forms, such as Gini impurity and entropy. With

Gini impurity, a Gini index can be calculated to reflect the average cumulative decrease attributed to a certain variable, then the Gini indices are ranked for feature importance. A third approach is to use deep learning algorithms which can detect complex dependence among the variables, there are techniques such as greedy search and averaged input gradient to obtain feature importance from the dependences (Wojtas and Chen, 2020).

1.2.3 Overfitting

The third issue of ML is overfitting. A ML algorithm is trained in a sub-sample of the data aiming to achieve prediction accuracy, but while doing so, it can pick up the noise and irrelevant information in the sub-sample. Overfitting is the situation when the model learns the sub-sample too well and relies on the irrelevant information for prediction, it can fail to generate to unseen data. ML algorithms can learn very specific (e.g., nonlinear) relationships in the sub-sample, thus they are susceptible to overfitting. In ML, limiting overfitting is very important. ML algorithms mainly use regularization and train-validation-test split to avoid overfitting (Storm et al., 2020).

The idea of regularization is to discourage complex models by putting a penalty when the model becomes too complex, therefore preventing the model from fitting very specific patterns in the sub-sample. For instance, in linear regression, LASSO regression and Ridge regression are the two prevailing approaches. In neural networks, one can decompose the final layer of the fitted neural networks into uncorrelated sub-networks to enabling the implicit regularization mechanism (Olson et al., 2018).

Train-validation-test split is to split the data into three sub-samples, one for training, one for validation, and one for testing. The model is trained on the training set, then its performance is validated and optimized using the validation set, and the model performance is finally evaluated

using the test set. When datasets are large, the train-validation-test split approach can be easily implemented. On smaller datasets, an alternative is to use k-fold cross validation.

1.4 Conclusion

The above sections discuss the strength and weakness of ML methods, with an emphasis on the comparison with econometric methods. ML methods and econometric methods have different objectives, but we argue that ML can enrich the economists' toolbox by complementing existing econometric methods and also bringing in novel insights. ML can increase the flexibility in data options, variables and functional form selection; it can bypass some limitations of the traditional econometric methods; combining ML methods with econometric methods can expand the boundaries of economic research.

Though ML methods require little domain knowledge to train a machine, economic theories can help researchers to choose the model family that can best describe the relations among variables. Economic theories are also essential in validating and explaining the findings of a trained model, hence can help to make the model scientifically interpretable.

Just as Leo Breiman "predicted" (Breiman, 2001), there is a shift from traditional statistical methods to ML in a wide range of science subjects. Guido Imbens, the 2021 Nobel laureate in economics, writes about ML methods that economists should know about and suggests that ML methods should be included in the core graduate econometrics sequences (Athey and Imbens, 2019). The authors conclude that "being familiar with these methods will allow researchers to do more sophisticated empirical work, and to communicate more effectively with researchers in other fields".

Following this trend, this paper uses data-driven ML methods to study two empirical questions. Chapter 2 will look at a novel deep learning framework to uncover the causal patterns

in credit default swaps (CDS) during the eurozone crisis in 2010. Chapter 3 will use two ML methods to establish the early warning system for financial crisis. Our datasets are structured medium size data, therefore traditional econometric methods can be applied as well, thus allowing the comparison between ML methods and econometric methods.

REFERENCES

Athey, Susan (2015): Machine learning and causal inference for policy evaluation. In *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Athey, Susan (2017): Beyond prediction: using big data for policy problems. In *Science* 355 (6324), pp. 483–485.

Athey, Susan (2018): The impact of machine learning on economics.

Athey, Susan; Imbens, Guido (2017): The state of applied econometrics: Causality and policy evaluation. In *Journal of Economic Perspectives* 31 (2), pp. 3–32.

Athey, Susan; Imbens, Guido (2019): Machine learning methods economists should know about.

Athey, Susan; Imbens, Guido; Wager, Stefan (2018): Approximate residual balancing: debiased inference of average treatment effects in high dimensions. In *J. R. Stat. Soc. B* 80 (4), pp. 597–623.

Belloni, Alexandre; Chernozhukov, Victor; Hansen, Christian (2014): High-dimensional methods and inference on structural and treatment effects. In *Journal of Economic Perspectives* 28 (2), pp. 29–50.

Breiman, Leo (2001): Statistical modeling: The two cultures. In *Statistical science* 16 (3), pp. 199–231.

Cai, Zongwu; Hong, Yongmiao (2003): Nonparametric methods in continuous-time finance: A selective review

Caiafa, Cesar F.; Sun, Zhe; Tanaka, Toshihisa; Marti-Puig, Pere; Solé-Casals, Jordi (2021): Machine learning methods with noisy, incomplete or small datasets. In *Applied Sciences* 11 (9), p. 4132.

Charpentier, Arthur; Elie, Romuald; Remlinger, Carl (2020): Reinforcement learning in economics and finance.

Charpentier, Arthur; Flachaire, Emmanuel; Ly, Antoine (2019): Econometrics and machine learning. In *Ecostat* 505 (505d), pp. 147–169.

Christiano, Lawrence J.; Eichenbaum, Martin S.; Trabandt, Mathias (2018): On DSGE models. In *Journal of Economic Perspectives* 32 (3), pp. 113–140.

E, Weinan (2021): The dawning of a new era in applied mathematics. In *Notices Amer. Math. Soc.* 68 (04), p. 1.

Efron, Bradley; Hastie, Trevor (2016): Computer age statistical inference. Algorithms, evidence, and data science; Cambridge University Press.

Einav, Liran; Levin, Jonathan (2014): Economics in the age of big data. In *Science* 346 (6210), p. 1243089.

Farrell, Max H.; Liang, Tengyuan; Misra, Sanjog (2021): Deep neural networks for estimation and inference. In *ECTA* 89 (1), pp. 181–213.

Ghoddusi, Hamed; Creamer, Germán G.; Rafizadeh, Nima (2019): Machine learning in energy economics and finance: A review. In *Energy Economics* 81, pp. 709–727.

Glymour, Clark; Zhang, Kun; Spirtes, Peter (2019): Review of causal discovery methods based on graphical models. In *Front. Genet.* 10, p. 524.

Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron (2016): Deep learning. Cambridge, Massachusetts: The MIT Press.

Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil et al. (2014): Generative adversarial nets. In *Advances in Neural Information Processing Systems* 27.

Gu, Shihao; Kelly, Bryan; Xiu, Dacheng (2020): Empirical asset pricing via machine learning. In *The Review of Financial Studies* 33 (5), pp. 2223–2273.

Gupta, Shivani; Gupta, Atul (2019): Dealing with noise problem in machine learning data-sets: A systematic review. In *Procedia Computer Science* 161, pp. 466–474.

Harding, Matthew; Hersh, Jonathan (2018): Big data in economics. In *IZA world of labor*.

Hastie, Trevor.; Friedman, Jerome H.; Tibshirani, Robert. (2009): The elements of statistical learning. Data mining, inference, and prediction. Cham: Springer International Publishing.

Hofman, Jake M.; Watts, Duncan J.; Athey, Susan; Garip, Filiz; Griffiths, Thomas L.; Kleinberg, Jon et al. (2021): Integrating explanation and prediction in computational social science. In *Nature* 595 (7866), pp. 181–188.

Hong, Shangzhi; Lynn, Henry S. (2020): Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. In *BMC Medical Research Methodology* 2020 (1), p. 199.

Hong, Yongmiao (2007): The status, roles and limitations of econometrics. In *Economic Research Journal* (5), pp. 139–153.

Hong, Yongmiao (2020): Foundations of modern econometrics. A unified approach. New Jersey: World Scientific.

Hong, Yongmiao (2021): Understanding modern econometrics. In *China Journal of Econometrics* 1 (2), p. 266.

Hong, Yongmiao; Wang, Shouyang (2021): Big data, machine learning and statistics: challenges and opportunities. In *China Journal of Econometrics* 1 (1), p. 17.

Kay H. Brodersen; Fabian Gallusser; Jim Koehler; Nicolas Remy; Steven L. Scott (2015): Inferring causal impact using Bayesian structural time-series models. In *The annals of applied statistics* 9 (1), pp. 247–274.

Kleinberg, Jon; Ludwig, Jens; Mullainathan, Sendhil; Obermeyer, Ziad (2015): Prediction policy problems. In *The American economic review* 105 (5), pp. 491–495.

LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015): Deep learning. In *Nature* 521 (7553), pp. 436–444.

Low, Hamish; Meghir, Costas (2017): The use of structural models in econometrics. In *Journal of Economic Perspectives* 31 (2), pp. 33–58.

Marsland, Stephen (2015): Machine learning. An algorithmic perspective. Second edition. Boca Raton, FL, London, CRC Press.

Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012): Foundations of machine learning. The MIT Press.

Molnar, Christoph (2022): Interpretable machine learning. A guide for making black box models explainable. Second edition.

Moraffah, Raha; Karami, Mansooreh; Guo, Ruocheng; Raglin, Adrienne; Liu, Huan (2020): Causal interpretability for machine learning - problems, methods and evaluation. In *SIGKDD Exploring. Newsletters.* 22 (1), pp. 18–33.

Mullainathan, Sendhil; Spiess, Jann (2017): Machine learning: An applied econometric approach. In *Journal of Economic Perspectives* 31 (2), pp. 87–106.

Olson, Matthew; Wyner, Abraham; Berk, Richard (2018): Modern neural networks generalize on small data Sets. In *Advances in Neural Information Processing Systems* 31.

Racine, Jeffrey S. (2008): Nonparametric econometrics: A Primer. In *FNT in Econometrics* 3 (1), pp. 1–88.

Russell, Stuart J.; Norvig, Peter (2021): Artificial intelligence. A modern approach. Fourth edition. Hoboken: Pearson.

Sarle, Warren S. (Ed.) (1994): Neural networks and statistical models. Proceedings of the Nineteenth Annual SAS Users Group International Conference.

Schmidhuber, Jürgen (2015): Deep learning in neural networks: an overview. In *Neural networks: the official journal of the International Neural Network Society* 61, pp. 85–117.

Solow, Robert M. (2010): Prepared statement on "Building a science of economics for the real world".

Solow, Robert M. (1957): Technical change and the aggregate production function. In *The Review of Economics and Statistics* 39 (3), p. 312.

Stetter, Christian; Mennig, Philipp; Sauer, Johannes (2022): Using machine learning to identify heterogeneous impacts of Agri-environment schemes in the EU: A Case Study. In *European Review of Agricultural Economics*.

Stiglitz, Joseph E. (2018): Where modern macroeconomics went wrong. In *Oxford Review of Economic Policy* 34 (1-2), pp. 70–106.

Storm, Hugo; Baylis, Kathy; Heckelei, Thomas (2020): Machine learning in agricultural and applied economics. In *European Review of Agricultural Economics* 47 (3), pp. 849–892.

Varian, Hal R. (2014): Big data: new tricks for econometrics. In *Journal of Economic Perspectives* 28 (2), pp. 3–28.

Varian, Hal R. (2018).: Artificial intelligence, economics, and industrial organization.

Wager, Stefan; Athey, Susan (2018): Estimation and inference of heterogeneous treatment effects using Random forests. In *Journal of the American Statistical Association* 113 (523), pp. 1228–1242.

White, Halbert (1988): Economic prediction using neural networks: the case of IBM daily stock returns. In: IEEE 1988 International conference on neural networks, 451-458 vol.2.

Wojtas, Maksymilian; Chen, Ke (2020): Feature importance ranking for deep learning. In *Advances in Neural Information Processing Systems* 33, pp. 5105–5114.

Zhang, Ying; Ling, Chen (2018): A strategy to apply machine learning to small datasets in materials science. In *NPJ Computatoinal Materials* 4 (1), pp. 1–8.

CHAPTER 2

CONTAGION AND SPILLOVERS DURING THE EUROZONE CRISIS: AN EMPIRICAL STUDY USING CONVOLUTIONAL NEURAL NETWORKS

2.1 Introduction

The eurozone crisis that started in Greece since 2010 had the most substantial negative impact on Europe's politics and economies ever since the World War II. The crisis quickly spread to Ireland and Portugal, then Spain, Italy and Cyprus. Five out of the 6 countries had to seek for external bailout because of their sovereigns' unsustainable fiscal condition. Italy, the only exception, was close to the brink and had the fourth highest public debt in the world in 2013, at 133 percent of its GDP[5]. During the turmoil, not only the periphery countries were caught in fire, core countries such as Austria, Belgium and France were also enduring high government bond yields and volatility. The eurozone crisis resulted in significant economic and social costs to the whole European Union.

Each of the aforementioned 6 countries had its unique crisis features. For Greece, the root of the crisis was the persistent fiscal irresponsibility and lack of reform in its government. For Ireland and Spain, it started with the housing bubble, which led to the banking crises, and subsequently their governments went down while trying to save the banks. Portugal and Italy struggled with a decade of loss of productivity and competitiveness, ending up with large public debts and current account deficits. Cyprus, as a small economy, endured tremendous loss

[5] IMF, World Economic Outlook (April 2016).

because the Cypriot banks hold too much Greek assets and failed to withstand the waves of the Greek sovereign crisis[6].

This chapter attempts to study the contagion and spillovers during the 2010–2013 eurozone crises using a novel machine learning (ML) approach. Why is it important to study the contagion and spillovers in the eurozone? The question can be answered threefold. First, it is the first time in modern history to witness sovereign crises breaking out consecutively in so many countries within a monetary union. Under a single currency, member states of the eurozone are unable to shield the economies using independent monetary policies like a standalone country. It is important to investigate the contagion and spillovers among different economies to facilitate structural reform and policy implementation in the eurozone and the broader European Union. Second, the eurozone crisis has great impact on the regional economic and monetary union in Europe, which also affects the globalization and economic integration of other regions. The depth of regional cooperation should be examined carefully to avoid possible channels of crisis transmission. Third, if contagion and spillovers exist, understanding the source of such linkage among crisis countries can provide valuable lessons for crisis management and macro-prudential regulation.

To study the contagion and spillovers during the eurozone crisis, there are four critical questions that need to be answered. First, are there contagion and spillovers during the eurozone crisis? Second, if the contagion and spillovers exist, is Greece the origin of all crises in the eurozone and how do the crisis countries affect each other? Third, are there contagion and spillovers from the crisis countries to the non-crisis countries inside the eurozone? Fourth, are there contagion and spillovers from the eurozone countries to other countries in the European Union?

These four questions are the core research topics in the existing literature on eurozone sovereign crises, but the answers differ to a great extent. Some studies find no evidence that

[6] See Baldwin and Giavazzi (2015) for a broad overview of the eurozone crisis.

Greece is the main crisis contributor (e.g., Bhanot et al., 2012; Caporin et al., 2018), while others attribute the regional disturbance fully to Greece (e.g., Arghyrou and Kontonikas, 2012; Missio and Watzka, 2011). Three factors may contribute to the divergence in the empirical findings. The first is data source. Some studies use only credit default swaps (CDS) data, some use CDS and country specific variables, and others use CDS and sovereign bond market data. The second is the different modeling strategies. The commonly used models include the vector autoregression and dynamic factor model. The third is the different definitions of contagion and spillovers used by the researchers.

Understanding the difference in the definitions of contagion and spillovers is fundamental to this line of work. Almost every study on contagion and spillovers starts with a definition (so does this one), and the definitions are usually defined to fit the specific modeling strategy of the study, contributing to the differences in empirical results. Existing literature has multiple meanings for contagion and spillovers. Here are three definitions from three seminal papers. Masson (1998) denotes contagion in the context of multiple equilibria, "changes in expectations that are not related to changes in a country's macroeconomic fundamentals" can lead to crisis in a country, the contagion is operated through the change of expectations. Spillovers refer to the interdependence among countries, especially the linkage of macroeconomic fundamentals. Dornbusch (2000) defines contagion as "spread of market disturbances—mostly on the downside—from one country to the other, a process observed through co-movements in exchange rates, stock prices, sovereign spreads, and capital flows." Spillovers means the co-movements among markets; it may be a result from contagion or from normal interdependence. Kaminsky et al. (2003) define contagion "as an episode in which there are significant immediate effects in a number of countries following an event", spillovers are the gradual cases that have "protracted effects that may cumulatively have major economic consequences." These authors' definitions vary and there is hardly a consensus (see Rigobon (2019) for a review).

Specifically, for the eurozone crisis literature, Broto and Pérez-Quirós (2015) document the different definitions and categorize them into three groups. The first group identifies the increased bivariate correlation between country pairs during the stress periods as contagion. The second group, after controlling for common fundamentals, identifies large shock transmission detection during the stress period as contagion. The third group usually identifies lagged volatility spillovers or Granger causality as contagion. Broto and Pérez-Quirós (2015) specify the difference between contagion and spillovers such that "spillovers, which is the lagged transmission of a shock, whereas contagion is simultaneous in nature." This difference between contagion and spillovers is consistent with Kaminsky et al. (2003) and many others (Mink and Haan, 2013; Bruyckere et al., 2013; Claeys and Vašíček, 2014).

In this study, our definition is in line with Kaminsky et al. (2003) and Broto and Pérez-Quirós (2015), the term "contagion and spillovers" includes both simultaneous contagion and lagged spillovers, observed through co-movements in the CDS market, characterized as pairwise cause-effect relationships. While this definition is consistent with previous literature, it is also drawn from our analytical model, which is a deep learning method called Temporal Causal Discovery Framework (TCDF).

The results of the TCDF provide answers to the four critical questions mentioned earlier. First, there are contagion and spillovers during the eurozone crisis. Second, Greece is by no means the black sheep of all the eurozone sovereign crises. Greece has a great impact on Cyprus, causing its banking and sovereign crisis, but Greece is not the cause for Ireland, Italy, Portugal and Spain. There is a chain of cause and effect from Ireland to Italy, then to Spain, and there is a confounding factor between Ireland and Portugal. Third, there are spillovers from the crisis countries to the non-crisis countries. Fourth, there is evidence of spillovers from the crisis countries to European Union member state outside of the eurozone.

This study contributes to the existing literature in three ways. First, to the author's knowledge, this is among the first studies that use ML methods to investigate the contagion and

spillovers during the eurozone crisis. Second, with the advance of causal machines in the ML community, this paper goes beyond ML's ability to solve the prediction problem. We use novel causal structure learning in economic research. Third, by adopting the TCDF, we are able to analyze the contagion and spillovers at a larger scale and in higher dimension where the results provide richer information than previous studies. Our results not only agree with some prior works, but also add new findings to the repository, which also confirm the current consensus on the causes of the eurozone crisis.

Before going into the details, we want to specify some terms used in this study. We use "crisis countries" to refer to Cyprus, Greece, Ireland, Portugal, and Spain. These five countries joined bailout programs provided by outside parties, such as the International Monetary Fund (IMF), European Stability Mechanism. The bailout programs signify their domestic sovereign crises during the period from May 2010 to March 2013.

"Periphery eurozone countries" refers to Cyprus. Greece, Ireland, Italy, Portugal and Spain. This definition reflects their political and geographical status in the eurozone (Bartlett and Prica, 2016). Italy did not receive a bailout program, so it is not in the crisis country group. The six "periphery eurozone countries" are the ones that suffer the most during the eurozone crisis (Baldwin et al., 2015).

Table 2.1. Country group in the EU in 2016, adapted from Bartlett and Prica (2016)

Country group	**Countries**
Core countries within the eurozone and the EU (Inner Core)	Austria, Belgium, Finland, France, Germany, Netherlands
Core countries outside the eurozone, within the EU (Outer Core)	Czech Republic, Denmark, Estonia, Latvia, Lithuania, Poland, Slovakia, Sweden, the United Kingdom
Periphery countries within the eurozone and the EU (periphery eurozone countries)	Cyprus, Greece, Ireland, Italy, Portugal, Spain,
Crisis countries	Cyprus, Greece, Ireland, Portugal, Spain

"Core countries" are in the inner circle of the European continent, they are also the ones that receive lesser impact during the eurozone crisis. The inner core and outer core are divided by their participation in the European Economic and Monetary Union. Table 2.1 shows the country groups within the EU.

This chapter proceeds as follows. Section 2.2 introduces the background of the eurozone crisis. Section 2.3 reviews the current literature on contagion and spillovers during the eurozone crisis. Section 2.4 presents the CDS data used in this paper. Section 2.5 describes the TCDF model and the vector autoregression model. Section 2.6 presents the results of the models. Section 2.7 discusses the findings. Section 2.8 concludes.

2.2 Background

This section intends to provide an overview of the eurozone crisis. There are two parts. First, we discuss the institutional setup of the eurozone and the buildup of imbalances in each country. Second, we try to identify the causes of the crises and discern their distinct nature.

2.2.1 The root and buildup of the crises

In retrospect, just like every crisis, the root of the eurozone crisis is the disorderly unwinding of economic imbalances (Wanna et al., 2015). The imbalances lead to insupportable public and private debts and foreign lending, and eventually, to crisis. The eurozone countries who suffered the most are the ones who borrowed the most (Baldwin et al., 2015). The source of such imbalances can be traced back to the deficient design of the European Economic and Monetary Union (EMU). A major reason for the imbalances' buildup is that the imbalances are conceived as a positive sign of financial market integration by the EMU (Peet and La Guardia, 2014).

Inside the eurozone, member states surrender the flexibility to wield their own monetary policy, but they are not compensated with sufficient protection against financial crisis from the European Union (EU). EU has insufficient institutions for monetary integration and it is lack of a response mechanism (Frankel, 2015). For example, the no-bailout clause (Article 125 TFEU[7]) forbids the European Central Bank (ECB) to lend money to member state in times of sovereign insolvency. The structural insufficiency of the EU, coupled with the favorable credit condition before the 2008 subprime crisis, are the main reasons for the large imbalances.

Going back in time, the EU and the eurozone all started in 1992 with the Maastricht Treaty, which promised an integrated and cooperative Europe for the European countries. The Treaty put forward the convergence criteria[8] for those planning to join the EU, among the criteria was the famous fiscal requirement of a less than 3% government deficit to GDP ratio and a less than 60% government debt to GDP ratio. With more countries meeting the criteria, the Stability and Growth Pact was signed in 1997 in the hope for an optimum currency area. EMU began in 1999 with 11 countries, and the euro was launched in 2002[9].

The creation of the single currency brought faith in stability and growth, and led to convergence of interest rate among members of the eurozone. Figure. 2.1 shows the interest rate of 10-year government bonds for European countries (except Cyprus) from 1990 to 2007. Before 1999, large heterogeneities existed among the countries, the market asked high returns for countries with weaker fiscal fundamentals, such as Italy, Spain and Portugal. Core countries like Germany, France and Belgium paid lower rate, reflecting promising market prospect. Since 1999, heterogeneities in macroeconomic status and fiscal policies of different member states were projected into a homogenous interest rate. Italy, whose rate was 14% in 1990, dropped to 3% in 2001. German rate dropped from 8% to 4%. All countries could access the market at a low rate since 2001.

[7] http://data.europa.eu/eli/treaty/tfeu_2016/art_125/oj
[8] https://www.ecb.europa.eu/ecb/orga/escb/html/convergence-criteria.en.html
[9] https://europa.eu/european-union/about-eu/euro/history-and-purpose-euro_en

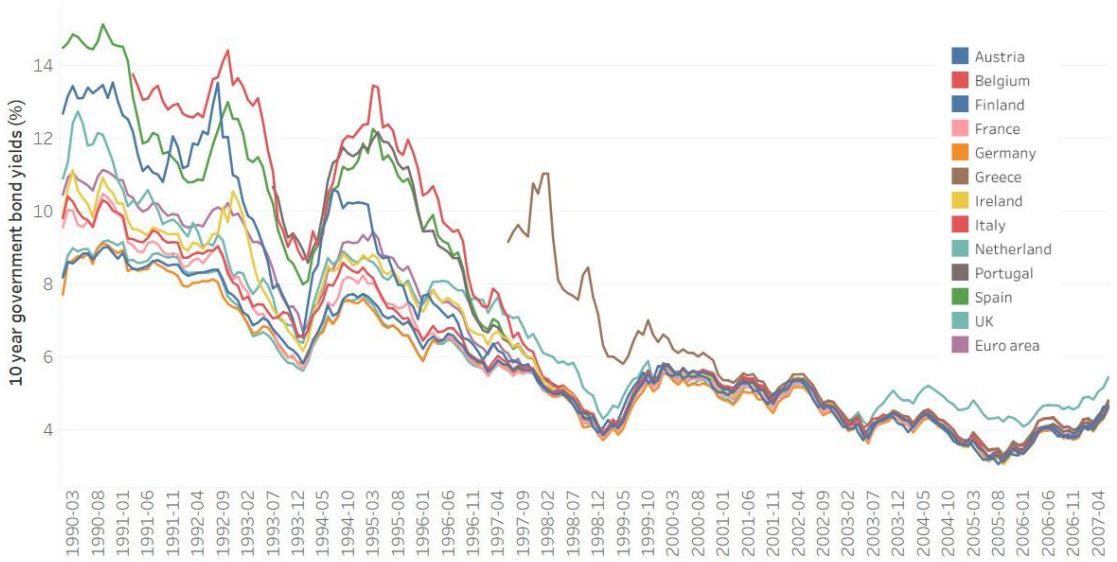

Figure 2.1. 10-year government bond yields (%) of European countries, source: OECD.stat.

Besides the introduction of the euro, another reason for the interest rate convergence is the "global savings glut", a term coined by Ben Bernanke. It described the situation in which the savings from developing countries (mainly China) were injected into the developed countries such as the U.S. and the eurozone countries[10]. With the global economic prosperity in the early 2000s, the periphery eurozone countries enjoyed several years of favorable credit conditions just like their core neighbors. In Figure 2.1, the only exception is the U.K., an outsider of the euro area. U.K.'s rate rose above the euro area average rate since 2003, which reflected the true market outlook of U.K.'s fiscal conditions. The unified interest rate between the periphery and the core countries covered up the divergence of their macroeconomic fundamentals.

In loose credit conditions, government tend to run sustained budget deficits that are financed by global investors (Cripps et al., 2011), the non-bank private sector becomes euphoric about the economic outlook and over-borrow because of the overly optimistic implicit signal about macroeconomic developments sent by the favorable credit conditions (McKinnon and Pill, 1998). Over-borrowing happens when the lending decisions of foreign financial institutions are

[10] https://www.federalreserve.gov/boarddocs/speeches/2005/200503102/

"guided by rough indicators of the country's macroeconomic performance and not by careful assessment of individual borrowers' abilities to repay" (Uribe, 2006). Over-borrowing in both the public and private sector of the periphery eurozone countries resulted in large imbalances. The imbalances would have been caught if they happened in a standalone country, but the prosperity of a unified monetary and economic region overshadowed the vulnerability of the member states.

There are three components of the imbalances. The first component is the over-accumulation of government debt. The Maastricht Treaty's threshold of a 60% debt to GDP ratio was clearly not enforced across the eurozone. Figure 2.2 shows the debt to GDP ratios of European countries from 1995 to 2007. The ratios rose above 60% in many countries, including France and Germany. Greece had an upward-sloping curve, with the ratio stood at 104% in 2007 and subsequently rose to 147.5% in 2010. On the left panel, Belgium had an abnormally high debt, which was on a par with Greece and Italy. Germany and France, rose above 60% since 2003, and their budget deficits also surpassed the 3% threshold set by the Maastricht Treaty. However, these two core members of the eurozone successfully evaded any official sanctions and revised the Safety and Growth Pact in favor of themselves in 2005 (Peet and La Guardia, 2014). Since then, the EU bodies no longer had the restraining power over sovereigns' fiscal conditions. Countries with weak fiscal fundamentals were spared from controlling their budget deficits and instituting necessary structural reforms. As a result, some eurozone countries had accumulated too much debt compared to their sustainable rate of productivity growth.

High level of government debts does not suggest an inevitable sovereign crisis. Countries like Belgium and Italy had a government debt of about 100% of their GDP, yet they stayed solvent throughout the eurozone crisis. Ireland and Spain, whose debt to GDP ratios were less than 40%, ended up with bailout programs (for Ireland and Spain, the housing bubble in the private sector had driven up the tax revenues, thus reducing the public debt, see Whelan (2014) for a review on Ireland).

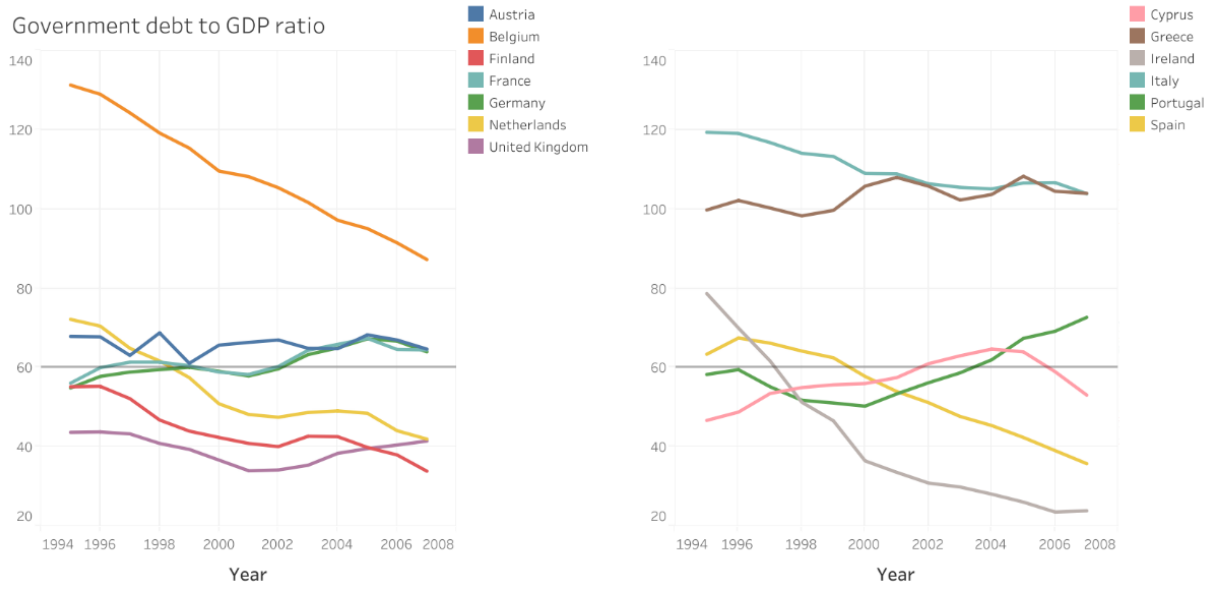

Figure 2.2. Government debt to GDP ratio. Left panel: non-crisis country. Right panel: periphery eurozone countries. Source: IMF Global Debt Database.

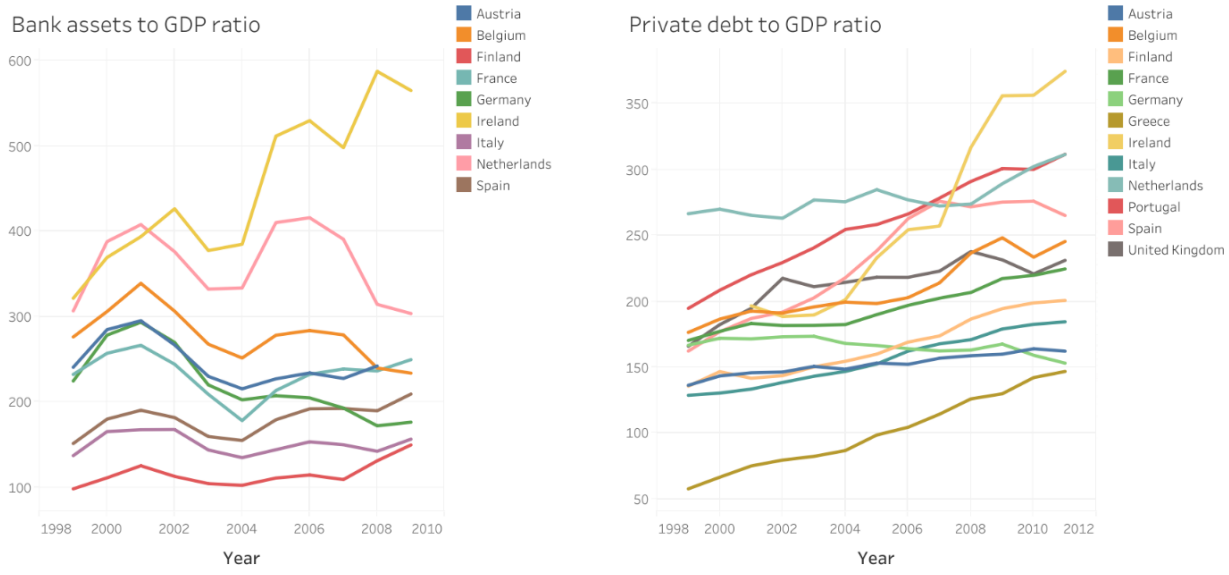

Figure 2.3 Left: Bank assets to GDP ratio in current prices from 1999 to 2009. Right: Private debt to GDP ratio in current prices from 1999 to 2011. Source: OECD.Stat.

The second component is the large private debts in almost all eurozone countries. In Figure 2.3, the left panel shows the banks' total asset to GDP ratio in current prices. Ireland stood out with an astounding percent at 588% of its GDP in 2008. Finland, who had the lowest ratio of all, had a stable and healthy percent around 120%. Italy and Spain were on better terms, following the same trend as most of their northern core neighbors. Due to data availability, Cyprus, Greece and Portugal are not included in the left panel.

The right panel of Figure 2.3 shows the private debt to GDP ratio in current prices. Ireland had the highest private debt among all countries, it reached a level of 374.5% in 2011, this is because of the pre-crisis housing bubble (Whelan, 2014). Spain, who also had a pre-crisis housing bubble, had the same issue with large private debt, it reached a percent of 265.3% in 2011. Portugal and the Netherlands had large private debts as well. At the bottom of the right panel, Greece had the smallest private debt to GDP ratio in all European countries.

The third component is the current account imbalances that built up inside the eurozone from 2000 to 2008. The current account deficits are a sign of danger for a balance of payment crisis in a standalone country. However, the European Commission never imagined the possibility of a balance of payment crisis inside the eurozone because of the single currency. The large current account deficits in the periphery eurozone countries were seen as "benign reflections of optimizing capital flows, instead of warning signals" (Frankel, 2015). Those current account imbalances were treated as harmless heterogeneity as in federal states in the U.S. However, the European Commission was not entitled to the equivalent administrative power and influence to its member states the same way as the U.S. had over its federal states. Besides the European Commission, other international organizations such as the OECD and the IMF also failed to send the warning signal of unsustainable trade imbalances and loss of competitiveness (Honkapohja, 2014). Eventually, the periphery eurozone countries built up large amounts of current account deficits in the early 2000s. If such imbalances are detected in a standalone

country, the large current account deficits would surely trigger the market response, as they are the signal of a loss of competitiveness.

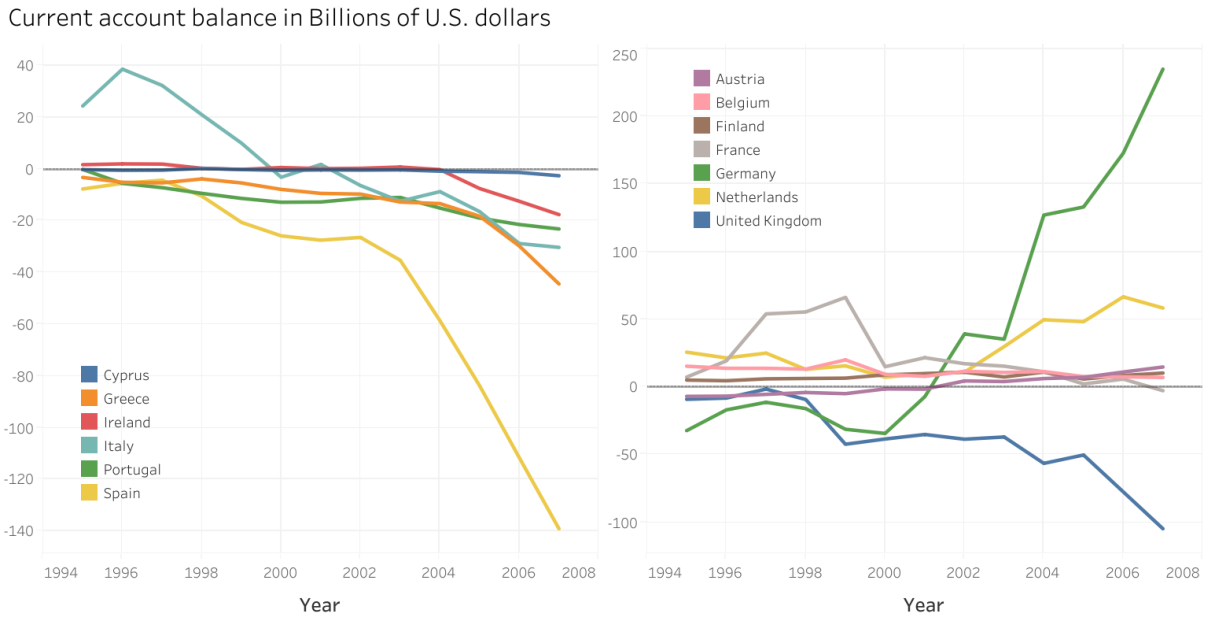

Figure 2.4. Current account balances in billions of U.S. dollars of the European countries from 1995 to 2007 in current prices. The left panel is the periphery eurozone countries, the right panel is the non-crisis countries. Source: IMF World Economic Outlook.

Figure 2.4 shows the current account balances in billions of U.S. dollars of the European countries from 1995 to 2007 in current prices. Periphery eurozone countries are on the left panel, non-crisis countries are on the right panel. Since 2000, Greece, Ireland, Italy, Portugal and Spain all dropped into large external deficits, while Cyprus seemed to maintain its balances. Spain had the largest external deficits in the eurozone, at 139 billion dollars in 2007. On the right panel, the U.K. ran an external deficit since 1998. However, given the size of its economy, the deficits only accounted for 3% of its GDP (see Figure 2.5). All other countries had external surplus.

Figure 2.5 shows the current account balances as a percent of GDP of the European countries. On the left panel, Greece had a large external deficit equaling to 14% of its GDP in 2007.

Portugal and Spain had a deficit around 10% of its GDP. Cyprus also built up an external deficit, which was 10.7% of its GDP. The size of the Cypriot economy was small, so the deficit was not obvious in levels in Figure 2.4. All six periphery eurozone countries had large current account deficits, five of them (Greece, Ireland, Portugal, Spain and Cyprus) had a sovereign crisis.

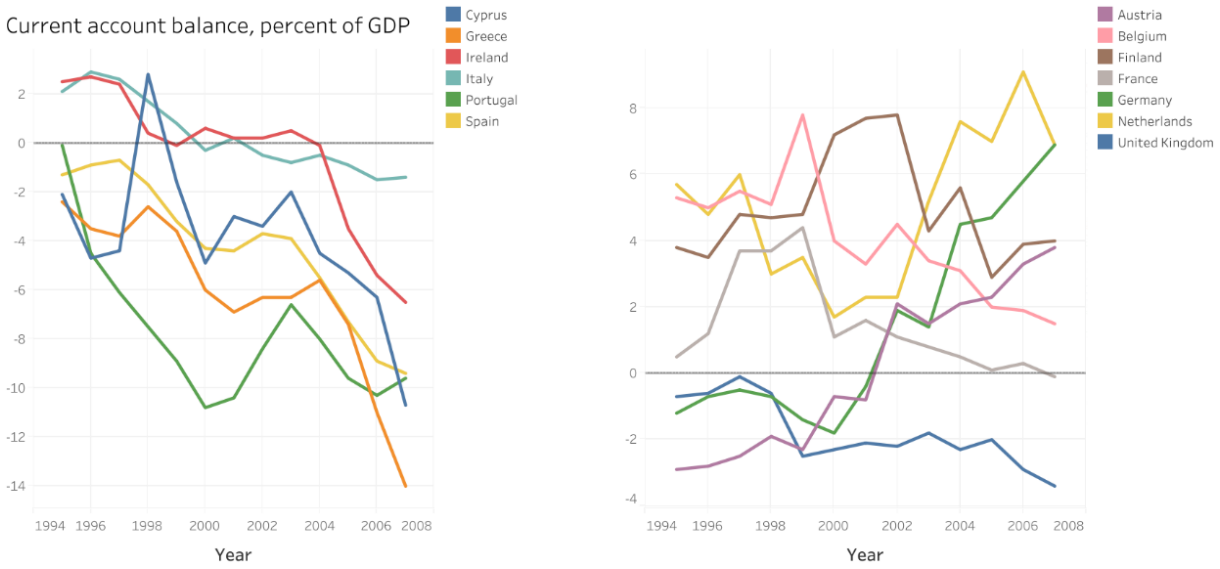

Figure 2.5. Current account balances as a percent of GDP of the European countries from 1995 to 2007. The left panel is the periphery eurozone countries, the right panel is the non-crisis countries. Source: IMF World Economic Outlook.

Figure 2.6 shows the current account balances in billions of dollars for the euro area and the EU from 1997 to 2016. The current account for the eurozone was above zero from 2001 to 2007, it dropped below zero in 2008 because of the subprime crisis, then it went up again in 2009 and kept growing since then. As a whole, the eurozone's current account was in balance before and after the crisis (Baldwin et al., 2015). The deficits countries had been importing from their rich neighbors inside the eurozone, mostly from Germany (see Figure 2.4). Hobza and Zeugner (2014) study the buildup of euro area imbalances and show that the current account deficits were linked through intro-eurozone financial inflows rather than trade flow, Germany was the

main driver but France played a crucial role for capital flows into the periphery countries as well.

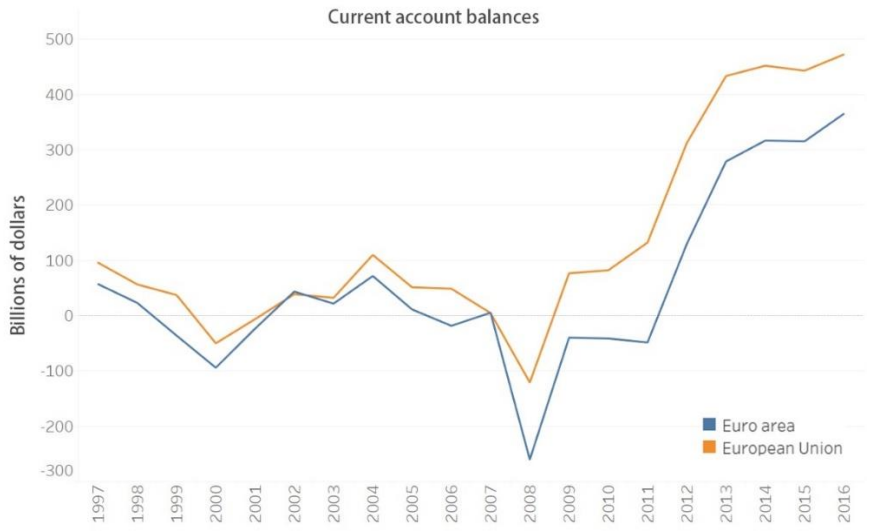

Figure 2.6. Current account balances of the euro area and the European Union from 1997 to 2016, in billions of dollars. Source: IMF World Economic Outlook.

In the periphery eurozone countries, the foreign capital inflows mostly went to the non-tradable sectors such as real estate and construction (Chen et al., 2013; Hobza and Zeugner, 2014), those sectors could not generate the growth and innovation that was essential to a country's productivity. What was even worse was that, the incoming investment drove up the housing price and factor prices, leading to a weak position in price competitiveness. The consequence of a decade of external deficits was the loss of productivity and competitiveness.

To sum up, the insufficient architecture and mismanagement of the EMU, combined with the loose credit condition before 2007, had allowed large internal and external deficits to build up in the periphery eurozone countries. The deficits were funded by foreign capital inflows (Baldwin et al., 2015; Croci et al., 2016). The EU bodies and the soon-to-be crisis countries turned a blind eye to the deteriorating fiscal fundamentals and loss of competitiveness, that eventually developed into a gap that was too wide to close.

2.2.2 The causes and nature of the crises

This section looks into the causes and the nature of the crises. First, we will look at the three causes for the crises. The first is the 2008-2009 global financial crisis; the second is the large banks and their close link to the sovereigns; the third is the loss of confidence in countries with high public debt. The insufficient and irresolute response of the EMU served as the catalyzer during the crises. Then, we will discuss the nature of crisis in each periphery eurozone country.

The first cause of the eurozone crises is the 2008-2009 global financial crisis. The global financial crisis triggered a series of shocks to the global capital market, causing sudden stop of cross border capital inflow all over the global. The six periphery eurozone countries all had either extensive public debts or private debts with a current account deficit, and foreign capital inflows were vital to sustain their debts. When sudden stops happened to the capital inflow, the imbalances could no longer sustain, triggering market response of a sharp decrease in bond prices (Constâncio, 2013). The periphery eurozone countries could not devalue their currency nor asked the central banks to bail out the government like a standalone country (Eichengreen and Gupta, 2018). Moreover, the eurozone countries did not have a lender of last resort, because the ECB was forbidden to bail out member states by EU Treaties. Grauwe and Ji (2013) points out that, when facing sudden stops, eurozone countries were more prone to self-fulfilling crisis than standalone countries.

Taking Greece as an example, the impact of the 2008-2009 global financial crisis on Greece was substantial. Two major Greek industries, shipping and tourism, fell victim to the global recession. The Greek economy contracted and so did the government tax revenue (Peet and La Guardia, 2014). Greece accumulated a 110.3% government debt to GDP ratio (highest in the eurozone), and a 7.1% government deficit to GDP ratio in 2008. Without foreign capital inflows, the huge public debts could collapse at any time. All the six periphery eurozone countries had

large debts funded by foreign capital inflows, the 2008 global recession revealed the high risk of such debts.

The second cause of the eurozone crisis is the large banks and their debt holdings of their home country's bonds. The European banks were too large before 2007. Veron (2007) argues that the banks in Europe had grown too large relative to its home country's GDP since the European integration. The "pan-European banks" spanned their services across the EU, but the bank supervision was only kept at the national level. The monetary policy was controlled by the EU-level bodies; however, no EU-level body had the authority to cope with large-scale bank failure. The author then warns that "faulty cross-border coordination and diverging national views could seriously hamper the ability of the authorities to respond speedily and effectively to an unfolding financial crisis." This was exactly what happened in Ireland. The Irish banks' total assets to GDP ratio was 588% in 2008 (see Figure 2.3), far exceeding other eurozone countries. When Irish banks were facing an imminent banking crisis, the sheer size of the banks required the Irish government to save them to avoid further economic disturbance, but the 588% ratio also suggested that the Irish government could not afford to save it.

The European banks tend to hold a relatively large share of sovereign debts of their home country. Table 2.2 shows a breakdown by sector of holdings of marketable debt for the European countries and the U.S. in 2007 and 2011 (from Merler and Pisani-Ferry, 2012). Before the subprime crisis, the U.K. and U.S. held -1.6% and 1.4% of their own debt in 2007. In contrast, the continental European banks had a much larger share of their home country's debts. German and Spanish banks held above 20%, Greece, Italy and Portugal held about 10%. Ireland was the only exception, with just 2.6%. Home bias is a measure that reflects to what extent a bank's share of domestic government debt exceeds the averages share of other countries' government debt (Horváth et al., 2015). Many studies have documented a large home bias in European banks' debt holding before the eurozone crisis (e.g., Horváth et al., 2015; Saka, 2020).

Table 2.2. Breakdown by sector of holdings of marketable debt, 2007 and 2011 (billions of national currency and, in parentheses, percent of total stock), modified from Merler and Pisani-Ferry (2012)

	Domestic banks	Central bank	Other public institutions	Other residents	Nonresident (excl. ECB)	Total
Greece	23.9 (10.6)	3.2 (1.4)	25.4 (11.3)	6.5 (2.9)	166.1 (73.8)	225.1
Ireland	0.8 (2.6)	n/a	0.1 (0.3)	1.2 (3.95)	28.8 (93.1)	30.9
Portugal	10.6 (9.1)	0.0 (0.0)	n/a	17.3 (15.0)	87.7 (75.9)	115.6
Italy	159.9 (12.1)	60.3 (4.6)	n/a	450.7 (34.2)	647.1 (49.1)	1317.9
Spain	74.3 (21.2)	9.2 (2.6)	26.5 (7.6)	73.3 (20.9)	166.7 (47.7)	349.9
Germany	456.9 (29.7)	4.4 (0.3)	0.5 (0.03)	317.1 (20.6)	761.5 (49.4)	1540.4
France	83.3 (13.0)	n/a	n/a	205.0 (32.0)	352.4 (55.0)	640.7
Netherlands	18.7 (8.9)	n/a	0.9 (0.4)	44.7 (21.4)	144.6 (69.2)	209.0
U.K.	-7.9 (-1.6)	2.4 (0.5)	0.8 (0.2)	337.3 (68.5)	160.2 (32.5)	492.8
U.S.	129.8 (1.4)	754.6 (8.2)	4616.5 (50.0)	1375.1 (14.9)	2353.2 (25.5)	9229.2

Source: Merler and Pisani-Ferry (2012)

This interdependence of banks' holding in home country's debt can create a "negative feedback loop" between the distressed banks and the distressed sovereign (Angelini et al., 2014; Saka, 2020). When the market loses faith in a sovereign's solvency, their banks, who have heavily invested in their own sovereign's bonds, are also caught in the fire. The underperforming banks depress the economy, thus depressed the already distressed sovereign's budget situation. The viscous cycle of deteriorating bank and sovereign can aggravate an ongoing crisis.

The third cause of the eurozone crisis is the loss of confidence in countries with high public debt and weak fundamentals. The market started to lose faith as the rating agencies (e.g., Moody, S&P) downgraded the sovereign bonds of the periphery eurozone countries, for example, S&P downgraded the Greek 10-year sovereign bond from A in 2009 to CCC in 2011. Many papers

study the influence of market assessments during the eurozone crisis. Santis (2012) shows a strong link between the rating downgrade in Greece and bond yields in other eurozone countries with weak fundamentals: Ireland, Portugal, Italy, Spain, Belgium and France. Grauwe and Ji (2013) find that the surges in the sovereign bond yields are "associated with negative self-fulfilling market sentiments that became very strong since the end of 2010".

The European Commission's inadequate crisis management also aggravated the market outlook. The leaders of the European Commission lacked the strength and the resolution to solve the Greek crisis. For example, France and Germany signed the Deauville Agreement amid the Greek crisis in October 2010 (in their own favor). This agreement served three purposes. First, it delayed sanctions on countries whose debt to GDP ratio was above 60%. Second, it required private-sector involvement in future bailouts, meaning larger debt write-downs for private investors. Third, countries amid a crisis would lose the right to vote in the EU Councils. Agreement like this symbolized the divides between the core countries and periphery countries. The discords in the eurozone exacerbated the economic outlook of the indebted countries and drove investors further away.

The above discussion concludes the three causes for the eurozone crisis. The following part will explain the nature and development of crisis in each country. Though the five crisis countries all received bailout loans, one should not consider the eurozone crisis as a pure sovereign debt crisis. Its nature is more akin to a balance of payments crisis (Higgins and Klitgaard, 2014). To help understand the crisis development, in Appendix A, Table A.2 shows a timeline of the eurozone crisis major events.

Greece had the highest level of public debt of all eurozone countries before the onset of crisis, its government debt to GDP ratio was 127.8% in 2009. On top of that, the political instability added fuel to the fire. In November 2009, the newly elected government accused their predecessor of "fabricating" the budget deficits, that the true deficits to GDP ratio was 12.7%,

almost twice as their predecessor's claim (Peet and La Guardia, 2014). This shocking revelation led to the downgrading of Greek bonds and soaring bond yields; the huge government debts could no longer be refinanced. The debt to GDP ratio amounted to 147.5% in 2010. Bond yield went from 6% in 2009 to an astounding 29% in 2012. The struck of the 2008-2009 global financial crisis, combined with the revelation of its true debt condition, placed Greece in a classical sovereign debt crisis. The government could no longer stay solvent on its own. At the early stage, the European Commission insisted on the "no bailout clause" and denied the possibility of asking IMF for help (Frankel 2015). These decisions inevitably exacerbated the situation. Speculations about a "Greek exit" emerged, under such pressure, in May 2010, the European Commission, ECB and the IMF (the three entities are named "the Troika"), issued a €110 billion euro bailout loan to Greece, which required the Greek government to implement austerity measures and structural reforms[11]. This marked the beginning of a series of crises in the eurozone.

Ireland, whose government debt was below the required 60% threshold, had a different story. Ireland had the largest banks' total asset to GDP ratio among all European countries (588% in 2008). It endured a housing bubble in the real estate market that was created by the loose credit conditions prior to the U.S. subprime crisis in 2008 (Whelan, 2014). When the bubble started to burst, the highly leveraged Irish banks started to encounter liquidity risk. The endangered banks were so large (pan-European) that the government must save them to avoid more severe consequences. The EU level authorities who had control over monetary policy failed to contain the emerging banking crisis (due to the lack of proper response mechanism, see Lane (2011) for a discussion). The Irish government, who was unable to maneuver any monetary policy, became insolvent themselves while saving the banks. Ireland had to receive its first bailout loan in

[11] See IMF website for bailout program details:
https://www.imf.org/en/News/Articles/2015/09/28/04/53/socar050210a

November 2010 (six months after Greece). Ireland only had a 42.4% government debt to GDP ratio in 2008, which was the second smallest of all eurozone countries. Ireland did not share the same fiscal distress as Greece. The nature of the Irish crisis was a banking crisis triggered by the global recession, then the banking crisis caused the sovereign crisis.

Portugal had long-term slow growth and loss of competitiveness coming into the crisis. It had built up large current account deficits (11.8% of its GDP in 2008). Portugal also had weak fiscal fundamentals, its government debts to GDP ratio had risen from 75.6% in 2008 to 100.2% in 2010. When its neighbors Greece and Ireland fell into the abyss of sovereign insolvency, concerns of the Portuguese sovereign had triggered sudden stops of the capital inflows. As the market sentiment of the Portuguese government bonds shifted, the Portuguese banks faced a backlash of holding too much home country bonds (9.1% of all Portuguese bonds were held by domestic banks). This had resulted in a typical negative feedback loop situation (Angelini et al., 2014). Portugal received a bailout loan in May 2011, and its situation was akin to a non-typical balance of payment crisis with sudden stops of capital inflows, joined by a negative feedback loop between the banks and the sovereign.

Spain had the highest level of current account deficits among all eurozone countries, measured at 145,274 billion dollars in 2008. The prolonged huge imbalance showed a loss of competitiveness of the Spanish economy (Baldwin et al., 2015). Similar to Ireland, Spain endured a housing bubble before the crisis, a large portion of banks' loans went to mortgage (33% of total loans in 2007) and construction (8% of total loans) that fueled the bubble (Quaglia and Royo, 2015). The banks also held too much Spanish sovereign bonds (21.1% of total Spanish bonds in 2007). When the sudden stops of capital inflows struck, the housing bubble started to burst. The negative market sentiment of the struggling neighboring countries also aggravated the economic outlook for Spain. The banks' large holding of the Spanish bonds also led to a negative feedback loop. The Spanish banks, received a bailout loan through the Spanish

government from the Troika in June 2012. The Spanish government stayed solvent throughout. Luckily, Spain had smaller banks, Spanish banks were of a similar size to core countries such as Germany and France, with a total asset to GDP ratio of 189.7% in 2008. It also had a healthy government debt to GDP ratio (39.7% in 2008). The Spanish banking crisis was not a fiscal matter; the culprit was the current account deficits (loss of competitiveness) and the banking sector.

Cyprus, the smallest one in the periphery eurozone countries (0.2% of the eurozone economy), was almost impossible to remain intact while all of its neighbors were experiencing an economic downturn. The Cyprus crisis started with the banking sector; Cypriot banks held excessive Greek assets. Zenios (2013) measures that the Cypriot banks face a "total exposure to Greek loans and sovereign debt worth 160% of GDP" in 2011. When the Greek bond was downgraded by rating agencies in 2010, the Cypriot banks were severely impacted. When the Greece installed the private sector involvement (PSI) in 2011, the Cypriot banks endured a loss equivalent to 23.03% of its GDP (Zenios, 2013). The burdens became too heavy to bear for the banks and the Cypriot government, so they sought help from the Troika. Cyprus was granted a bailout loan in March 2013. The Cyprus crisis was the collateral damage of its underperforming neighbors. Their banks' close linkage with the Greek economy served as the amplifier. Besides the banking sector, Cyprus had large private debts (118% of its GDP in 2010), and current account deficits (10.7% of GDP in 2010), but its public debts (55% of GDP in 2010) were well maintained, suggesting that the Cypriot crisis was certainly not a fiscal one.

Italy had never joined a bailout program, but it had its share of the crisis experience. Being the third largest economy inside the eurozone, Italy held a large government debt (119% of GDP in 2009). Italy's banking sector and the current account deficits were on a reasonable trajectory. But there is another layer of instability, which was Italy's political weakness. In November 2011, the leadership change in the Italian government "marked the economic and political crisis

in Italy" (Romano, 2021). Orsi (2013) explains that, given the adverse fiscal and political environment, "the Italian state went bankrupt in summer 2011." Though the Italian government never joined a bailout program, it benefited hugely from several EU programs. In August 2011, the ECB, started to purchase Italian and Spanish bond from the market. In December 2011, the ECB launched the Long Term Repo Operations (LTRO) mechanism, which injected lots of liquidity into the Italian financial market. These measures saved Italy from the fate of insolvency. Though Italy did not receive any bailout loans, Di Quirico (2010), Romano (2011) and other authors consider Italy as a crisis country, just as Greece.

In 2012, the European Commission and the ECB had taken measures to rescue the sinking euro area. The turning point was December 2012, when the ECB president Mario Draghi made the famous speech that "the ECB is ready to do whatever it takes to preserve the euro". The speech turned the negative market beliefs into positive. Since then, the collective actions of the crisis countries, EU authorities, and international organizations helped to save the eurozone countries.

From the above discussion, we can see that all the six periphery eurozone countries had large current account deficits, they all share the problem of a loss of productivity and competitiveness. The crises resemble a typical balance of payment crisis in a standalone country, which is unimaginable for countries in a currency union. It is natural to raise the question of how those crises link to each other? The next section gives a review of the existing literature on the contagion and spillovers during the eurozone crisis.

2.3 Related literatures

The primary focus of this paper is to use ML method to study the interdependences among the European countries during the eurozone sovereign crises. The most commonly used data to

study the sovereign crisis is the credit default swap (CDS) of sovereign bonds. This paper uses the daily spreads of CDS of the 5-year sovereign bond. The daily dataset can provide ample data points for a ML setup. Other than the CDS data, the existing literature also uses stock market price, government bond yield, and other macroeconomic data. In order to compare the empirical findings, this literature review only focuses on the studies that use the CDS data.

Previous studies mainly use two approaches to study the contagion and spillovers, one is to measure the contagion and spillovers among different countries (e.g., Alter and Beyer, 2014; Glover and Richards-Shubik, 2014). The other is to study the determinants of the CDS (e.g., Beirne and Fratzscher, 2013; Arghyrou and Kontonikas, 2012). Common methodologies include ordinary least squares (OLS), vector autoregression (VAR), event studies, autoregressive conditional heteroskedasticity model (ARCH), nonlinear regressions, etc. Their findings can also be put into two groups, by whether contagion and spillovers are detected.

The first group of researchers finds no contagion and spillovers. Caporin et al. (2018) use standard quantile regression and Bayesian quantile regression with heteroskedasticity to study the sovereign risk shift-contagion in major eurozone countries. They find almost no presence of shift-contagion in their sample periods. There is no correlation between risk spillover and the sign of the shock (Greek crisis), meaning that there is little contagion, even during the most intense crisis period. Bhanot et al. (2012) analyze the relations in the CDS spreads using ARCH that include time-varying volatilities and changes in fundamentals, their results show that there is no conditional correlation in the CDS spreads between Greece and PIIGS (Portugal, Ireland, Italy, and Spain). They find no evidence of contagion from Greece to PIIGS, nor to the eurozone core countries.

Glover and Richards-Shubik (2014) use the CDS data and macroeconomic data to estimate a network model of credit risk to measure market expectations of the spillovers of sovereign defaults. After the model is calibrated, it is used to conduct simulation for the short-run effect

of a default. Their results show only tiny spillovers to other sovereigns, that each $1 of debt directly lost in default is linked with an expected loss of 2 cents from additional defaults in other countries. Koutmos (2018) employs a VAR model to study the dynamic interdependencies of CDS spread among several EU countries between October 2004 and July 2016. The author discovers that the interdependencies vary across different periods of time, but there is no empirical evidence to show that Greece has transmitted the sovereign crisis to other countries. Beirne and Fratzscher (2013) use a comprehensive linear model to study the drivers of sovereign risk (CDS spread) for 31 countries during the eurozone crisis, they found that the country specific variables on macro fundamentals play the biggest role, whereas the regional contagion and spillovers are inconsequential, even for the eurozone countries.

The second group of researchers detects certain levels of contagion and spillovers during the eurozone crisis. Arghyrou and Kontonikas (2012) apply a convergence-trade model to study the pricing behaviors of CDS spreads. They discover that most eurozone countries have experienced contagion from Greece during the crisis period. Bampinas et al. (2020) investigate both the sovereign bond market and the CDS market to study cross-border and intra-market linkage in the eurozone countries from 2006 to 2018, they adopt the excess correlation concept of Bekaert et al. (2005) and use the local Gaussian correlation approach to study the contagion. Their results show that contagion occurs during different periods, in particular, Italian and Spanish CDS spreads spill towards all European CDS spreads around November 2011.

Broto and Pérez-Quirós (2015) use a dynamic factor model to decompose the sovereign CDS spreads of ten OECD economies, they find three factors: a common factor, a factor driven by peripheral eurozone countries and a country specific factor. By utilizing the three factors in a novel methodology to characterize contagion, they discover that contagion has played a non-negligible role in the peripheral eurozone countries since the onset of the crisis. Buchholz and Tonzera (2016) use a multivariate GARCH model to study the sovereign CDS spreads of 17

countries from 2008 to 2012. They find strong evidence for both fundamentals and non-fundamentals based contagion among those countries.

Gómez-Puig and Sosvilla-Rivero (2016) use the logit model to study whether the sovereign risk is transmitted through "pure" or "fundamentals-based contagion" during the eurozone crisis. Their findings confirm the coexistence of "pure" and "fundamentals-based contagion". Kalbaska and Gątkowski (2012) study the CDS spreads of several European countries, they employ the EWMA correlation analysis and the Granger causality test, and find evidence of the contagion effect since August 2007. Missio and Watzka (2011) estimate a dynamic conditional correlation model to assess if contagion is identifiable during the eurozone crisis. Their findings confirm the existence of contagion in the euro area.

There are other papers who use the government bond dataset to study the contagion and spillovers between the bond market and CDS market, most of them find evidence supporting the existence of contagion (e.g., Claeys and Vašíček, 2014; Croci et al., 2016; Cronin et al., 2016).

Table 2.3 summarizes the related literatures. The second and third columns show the data and model used in the studies, the last four columns show their answers to the four questions. First, are there contagion and spillovers during the Eurozone crisis? Second, is Greece the origin of all the crises? Third, are there contagion and spillovers from the crisis countries to the non-crisis countries inside the eurozone? Fourth, are there contagion and spillovers from the eurozone countries to other countries in the European Union? The top panel shows the results of the first group of researchers, the bottom panel shows the results of the second group. The last row shows the results of this paper. Y stands for existence of contagion and spillovers, N stands for small or no contagion and spillovers, N/A means that this question has not been studied in the paper.

Table 2.3. Related studies on eurozone crisis contagion and spillovers

Paper	**Data**	**Model**	**Q1**	**Q2**	**Q3**	**Q4**
No contagion and spillovers						
Bhanot et al. (2012)	Bond spreads and CDS spreads, daily	Autoregressive conditional heteroskedasticity	N	N	N	N/A
Beirne and Fratzscher (2013)	Bond spreads, CDS spreads, and macroeconomic data, quarterly	Comprehensive linear model	N	N	N	N
Glover and Richards-Shubik (2014)	CDS spreads and macroeconomic data, quarterly	Financial network model	N	N	N	N
Caporin et al. (2018)	Bond spreads and CDS spreads, daily	Standard and Bayesian quantile regression	N	N	N	N/A
Koutmos (2018)	CDS spreads, weekly	Vector Autoregression	Y	N	Y	Y
Contagion and spillovers exist						
Missio and Watzka (2011)	Bond spreads, daily	Dynamic conditional correlation models	Y	Y	Y	N/A
Arghyrou and Kontonikas (2012)	Bond spreads and macroeconomic data, monthly	Convergence-trade model	Y	Y	Y	N/A
Kalbaska and Gątkowski (2012)	CDS spreads, weekly	Exponentially Weighted Moving Average and Vector Autoregression	Y	Y	N	N
Broto and Pérez-Quirós (2015)	CDS spreads, weekly	Dynamic factor model	Y	Y	N	N
Buchholz and Tonzera (2016)	CDS spreads and macroeconomic data, daily	Multivariate Generalized Autoregressive Conditional Heteroskedasticity	Y	Y	Y	Y
Gómez-Puig and Sosvilla (2016)	Bond spreads, daily	Logit model	Y	Y	Y	N/A
Bampinas et al. (2020)	Bond spreads and CDS spreads, daily	Bootstrap test using local Gaussian correlation	Y	N	Y	N/A
This paper	CDS spreads, daily	Temporal Causal Discovery Framework	Y	N	Y	Y

In Table 2.3, there is an obvious divergence of opinions between the two groups of researchers. For the second question of whether Greece has triggered the crisis in other sovereigns, it is not surprising to see that the authors could not reach a consensus. As explained in section 2.2, the periphery eurozone countries have areas of commonality while going into the crisis (the high external debts and weak fundamentals), but the nature of their crises differs. The difference in the empirical findings has three reasons. First, the datasets are different. Though all the studies use the CDS data, some use the 5-year CDS while others use the 10-year CDS. The frequencies vary from daily, weekly, monthly and quarterly. The CDS data can be the market closing price, or a composite index deduced from prices in several markets. Other than the CDS data, some studies also use macroeconomics data to control for country specific factors, some use the sovereign bond spreads and the CDS spreads to fit a model. Second, the modeling strategies are different, some use the model-driven approach to build structural models (e.g., Glover and Richards-Shubik, 2014; Arghyrou and Kontonikas, 2012), some use regression models (e.g., Caporin et al., 2018; Koutmos 2018; etc.). Third, there are nuances in the measurement of contagion and spillovers across the studies. Some use Granger causality, while some use their self-defined quantifiable measures.

All the researchers use traditional econometric methods to study the underlining linkage among European countries. This is partially because of data availability of the macro fundamental variables, most of them are monthly or quarterly data. The low frequency data has limited observations during the crisis periods, making it unsuitable for ML methods. Also, the research question on contagion and spillovers is causal, which is not the expertise of ML. For those reasons, the existing literature only has few applications of ML methods in the eurozone crisis, but their focus is on prediction, rather than contagion and spillovers.

With recent development of ML techniques in the causal discovery literature, one can employ the deep learning framework to discover the causal relationships in time series data. This paper employs the Temporal Causal Discovery Framework (TCDF), which uses the

convolutional neural network as the core algorithm, to study the causal relationships in the European CDS market.

2.4 Data description

It is a common practice to use the sovereign CDS data to study contagion and spillovers during sovereign crises, both in academia and the press (Augustin, 2014). This paper uses the daily 5-year sovereign CDS spreads in 13 European countries from 3 October 2005 to 31 December 2015, totaling country-daily 2,670 observations. The daily CDS data contains only weekday prices, excluding weekends. The CDS data is collected from Markit and Bloomberg.

Like most other studies, this paper uses the 5-year sovereign CDS spreads. The 5-year sovereign CDS is the most liquid among all maturities in the market, it has the largest number of transactions from which the daily CDS spreads can be deduced. The CDS spreads are based on the USD-denominated CDS contract (U.S. dollar is the standard currency in the CDS market).

The 13 countries include the six periphery eurozone countries: Cyprus Greece, Ireland, Italy, Portugal and Spain; six core eurozone countries: Austria, Belgium, Finland, France, Germany, the Netherlands; and one country outside the eurozone, the United Kingdom. Because the emphasis of this paper is to study the contagion inside the eurozone, and also due to data continuity, other EU countries such Denmark, Sweden and Norway are not included in the analysis.

There are some missing data for Greece in the period from March 2012 to June 2013. During this time, the Greek bond yields and default risk were so high that there were almost no CDS transactions in the market. The Markit CDS data was a composite price calculated from market prices, so there were not sufficient transactions to get CDS data for this period. To solve this problem, we use the Bloomberg CDS data on Greek sovereign CDS for the missing data points.

Though Bloomberg uses the mid-day quote of the CDS, while Markit uses a composite price, their Greek sovereign CDS data follows almost identical trends in other periods.

To observe the dynamic interdependencies between CDS spreads among the EU countries, we split the data into 4 phases: pre-crisis, crisis buildup, euro area recession and post-crisis. In our analytical models (TCDF and VAR), there is no time varying measure for contagion and spillovers, we cannot monitor the spillover dynamics using the whole sample, splitting the data allows us to study the pattern of contagion and spillovers in different crisis periods. Many other papers have also taken this approach (e.g., Kalbaska and Gątkowski, 2012; Koutmos, 2018).

The first phase is the pre-crisis period, with the start date of 3 October 2005. This is the earliest date of the data. Most of the literatures on subprime crisis use 9 August 2007 as the starting date for the subprime crisis (e.g., Longstaff, 2010). It is the day when BNP Paribas announced that it would freeze $2.2 billion worth of funds in the U.S. subprime mortgage market. Hence, we choose 31 July 2007 as the pre-crisis end date. During this period, worldwide economies enjoyed the prosperity of stable economic growth.

The second phase is the crisis buildup period. The subprime crisis is one of the three causes of the eurozone crisis. Starting from 2007, countries like Greece, Ireland and Spain endured sudden stops of foreign capital inflows, the credit crunch put pressure on the banking sector and the sovereigns. Therefore, we pick the start date of the subprime crisis as the starting date of the eurozone crisis buildup period, that is 9 August 2007. The end date of this period is the day before the first Greek bailout, that is 30 April 2010.

The third phase is the euro area recession period, the start date is 3 May 2010, when Greece received its first bailout loan. This bailout marked the beginning of the series of sovereign crises inside the eurozone. The end date of this recession period is 29 March 2013. We pick this date according to the European Commission's economic forecast in Spring 2013[12]. The start and end

[12] https://ec.europa.eu/economy_finance/publications/european_economy/2013/ee2_en.htmt

dates of the eurozone crisis are consistent with the OECD recession indicators for the eurozone[13].

The fourth phase is the post-crisis recovery period. The start date is 1 April 2013, following the last day of the recession period. The European Commission claimed that the euro area had restored its economic vitality back to the 2008 level by the end of the 2015[14], so we pick the end date of 31 December 2015.

Table 2.4. Crisis phases and data availability

Time	Phase	Data availability
3 October 2005 ~ 31 July 2007	Pre-crisis	12 countries (excl. U.K.), 417 observations.
1 August 2007 ~ 30 April 2010	Crisis buildup	12 countries (excl. U.K.), 718 observations
3 May 2010 ~ 29 March 2013	Euro area recession	13 countries. 760 observations.
1 April 2013 ~ 31 December 2015	Post-crisis	13 countries. 715 observations.

Our data partition is consistent with the studies that split the data into different crisis periods, but the exact periods vary across studies. For instance, Kalbaska and Gątkowski (2012) choose August 2007 as the start date of the eurozone crisis, while we consider it as the start of crisis buildup period (phase 2). To better compare our results with other studies, we analyze our data not just on phase 2, but also phase 2 and 3 together. Table 2.4 shows the 4 phases with the data availability. For phase 1 and 2, the CDS data of the U.K. is sparse, therefore the U.K. data is excluded from phase 1 and 2.

[13] The OECD recession indicator: https://fred.stlouisfed.org/series/EUROREC
[14] https://ec.europa.eu/commission/presscorner/detail/en/IP_17_2401

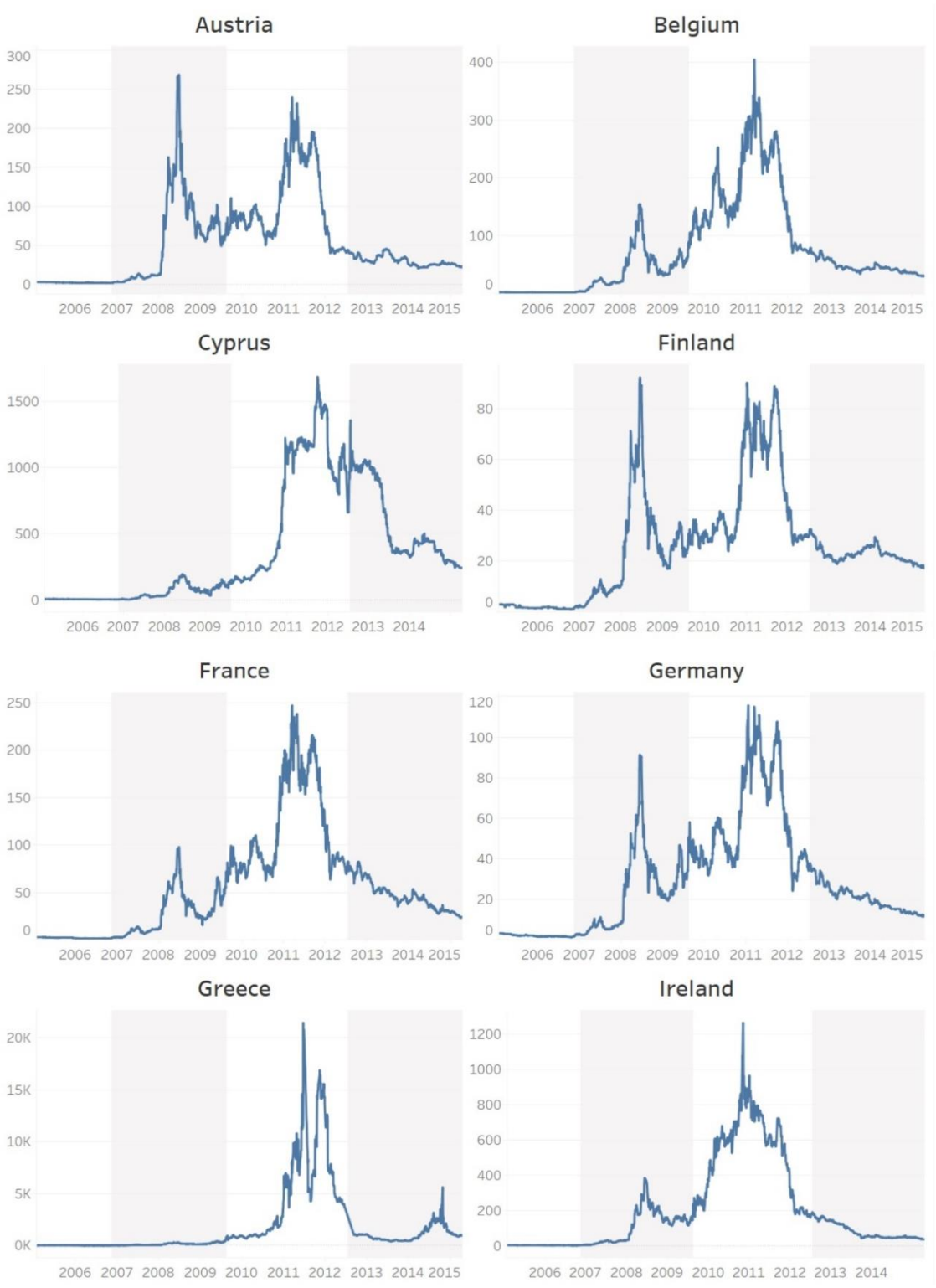

Figure 2.7. Time series plot of CDS spreads in basis points for Austria, Belgium, Cyprus, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal and Spain.

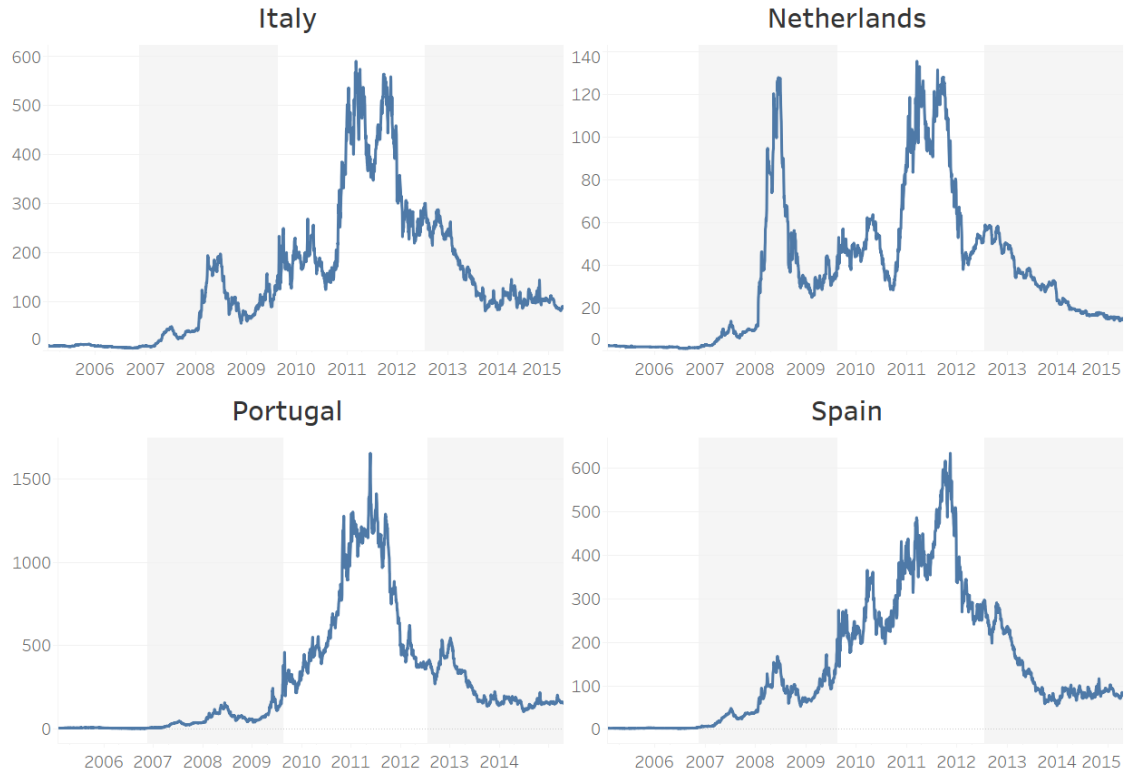

Figure 2.7 continued.

Figure 2.7 shows the time series plot of CDS spreads in basis points of the 12 eurozone countries from 3 October 2005 to 31 December 2015, the U.K. data is excluded due to data availability. The shaded background shows the 4 phases, respectively. In phase 1, all 12 countries' CDS spreads remain flat, reflecting the global economic growth and stability during the pre-crisis period. This tranquility in all countries suggests that there is no shock transmission during phase 1. In phase 2, the crisis buildup period encompasses the 2008-2009 global financial crisis. All 12 countries saw a rise in the CDS spreads. The rise in Cyprus, Greece and Portugal is not conspicuous in the figure, but if one looks at the scale on the y-axis, the Cypriot CDS spreads had reached as high as 195 in basis point, 145 for Portugal and 295 for Greece. For all the periphery eurozone countries, the rises in phase 3 dominate the rises in phase 2.

In Appendix A, Table A.2 shows the descriptive statistics of the CDS spreads in basis points. Table A.3 shows the correlation matrix of the log first differences of the CDS spreads. In those two tables, Panel A reports results for phase 1, Panel B reports phase 2, Panel C reports results for phase 3, Panel D reports phase 4, Panel E reports results of the full sample. In Table A.2, all the panels show that the periphery eurozone countries have the highest mean and median among the 13 countries, in particular, Greece leads with a large margin. During the pre-crisis periods (phase 1), the CDS spreads of the soon-to-be crisis countries are already higher than that of other countries. In retrospect, this suggests that the CDS spread is a potent measurement of market risk of the sovereign bonds.

The pairwise correlation among all the countries is near zero in phase 1, as shown in Table A.3 panel A, confirming the speculation in Figure 2.7 that there is no spillover during phase 1. Then, in phase 2, a significant increase in the correlation coefficients can be observed in panel B, which is an indicator for CDS spreads co-movement. This is unsurprising because all countries were affected by the 2008-2009 global financial crisis concurrently. In phase 3, the coefficient values start to show more variations. The correlation coefficient between Greece and Cyprus turns negative, which is the only negative coefficient in this panel. Greece, at the center of the vortex during phase 3, has an average pairwise correlation of 0.632 with other crisis countries. Its average correlation with non-crisis countries is 0.453, suggesting that the interdependency between Greece and non-crisis countries is lower than that of the crisis countries. In phase 4, the correlations numbers decrease in magnitude, implying stability during this period.

Augmented Dickey-Fuller (ADF) test is conducted on the CDS spreads data. We use the log-difference and log-level of the CDS spreads, the lag length is based on Akaike Information Criterion (AIC). The results are shown in Table 2.5 and Table 2.6. The log-difference of CDS

spreads is stationary, while the log-level of CDS spreads fail to reject the null hypothesis of a unit root. We use the stationary log-difference CDS spreads in our data analysis.

Table 2.5. Augmented Dickey-Fuller test statistics for the log-difference of CDS spreads

Country	Phase 1	Phase 2	Phase 3	Phase 4	Full sample
Austria	-5.498*	-5.440*	-6.975*	-7.842*	-11.656*
Belgium	-6.135*	-6.053*	-7.973*	-9.133*	-13.226*
Cyprus	-8.293*	-10.985*	-7.797*	-10.834*	-19.721*
Finland	-7.974*	-6.964*	-7.659*	-10.398*	-15.063*
France	-4.825*	-6.782*	-10.198*	-9.740*	-15.104*
Germany	-7.959*	-7.531*	-8.662*	-8.276*	-15.286*
Greece	-7.046*	-9.307*	-8.487*	-7.101*	-15.497*
Ireland	-6.900*	-6.324*	-10.021*	-9.060*	-15.248*
Italy	-3.230*	-6.660*	-11.366*	-6.025*	-14.624*
Netherlands	-6.527*	-7.576*	-9.081*	-8.795*	-15.051*
Portugal	-4.192*	-7.593*	-11.344*	-7.646*	-17.285*
Spain	-5.953*	-6.544*	-11.927*	-7.042*	-15.793*
UK			-9.817*	-8.218*	

Augmented Dickey-Fuller test statistics for the log-difference of CDS spreads. Lag length is based on the AIC. An asterisk (*) indicates statistical significance at least at 5% level to reject the null hypothesis of a unit root.

Table 2.6. Augmented Dickey-Fuller test statistics for the log-level of CDS spreads

Country	Phase 1	Phase 2	Phase 3	Phase 4	Full sample
Austria	-1.089	-1.089	-0.883	-0.139	-0.786
Belgium	-2.177	-2.177	-1.506	0.301	-1.104
Cyprus	-0.936	-0.936	-0.429	-0.355	0.219
Finland	-0.595	-0.595	-1.709	-0.123	-0.420
France	-0.854	-0.854	-2.384	0.085	-1.532
Germany	-0.313	-0.313	-0.177	-0.098	-0.498
Greece	-0.543	-0.543	-2.558	-3.571	-3.454

Ireland	-1.475	-1.475	-1.040	1.532	0.292
Italy	-0.717	-0.717	-2.699	0.369	-1.297
Netherlands	-1.495	-1.495	-1.682	0.257	-1.735
Portugal	-0.092	-0.092	-2.008	0.729	-1.380
Spain	-1.509	-1.509	-2.333	1.147	-1.067
UK			-0.955	-0.023	

Augmented Dickey-Fuller test statistics for the log-level of CDS spreads. Lag length is based on the AIC. All series fail to reject the null hypothesis of a unit root.

Some papers who study the CDS spreads data also use other explanatory variables, such as macro fundamental indicators, credit agency rating scores (e.g., Bampinas et al., 2020; Beirne and Fratzscher, 2013). Our paper follows Broto et el. (2015), Bartlett and Prica (2016), and others, who only use the CDS spreads time series data.

The reason for excluding the fundamental variables is due to data frequency. We use ML methods (TCDF) to study the CDS market. The ML method requires a relatively larger dataset. All relevant macroeconomic data is compiled monthly or quarterly, which offers too few data points during the crisis periods. Extrapolating a monthly series into a daily series is far from ideal. Also, the TCDF framework fits well with univariate time series data. Therefore, we only use the CDS spreads for data analysis.

2.5 Analytical frameworks

We use two methods for analyzing the contagion and spillovers during the eurozone crises. One is a ML method, the TCDF, the other is a traditional econometric method, Granger causality. By using both methods on the same data, we can compare the results and understanding the pros and cons of each method.

2.5.1 Temporal Causal Discovery Framework (TCDF)

Temporal Causal Discovery Framework (TCDF) is a novel ML algorithm developed by Nauta et al. (2019), it is among the first group of ML algorithms to incorporate deep learning into the causal discovery framework. TCDF uses Attention-based Dilated Depthwise Separable Temporal Convolutional Networks (AD-DSTCNs) to predict time series, then uses the attention scores obtained from the predictions to perform Attention Interpretation. After that, it applies Causal Validation and Delay Discovery to infer cause-effect relationships in the time series.

TCDF can be categorized as a deep learning algorithm in the ML methods, as its core algorithms are convolutional neural networks (CNN). At the same time, TCDF also belongs to the broader community of Causal Learning (CL) algorithms. CL is the recent development of combining ML and causal inference. Its goal is to reveal causal information by analyzing purely observational data. Causal learning goes "beyond machine learning due to its power of uncovering data generating processes" (Chen et al., 2022). We have briefly introduced some CL methods in section 1.3 (causal random forest developed by Wager and Athey (2018), etc.), here we will discuss the CL methods in more details.

There are two fundamental tasks in CL, one is to learn causal effects, and the other is to learn causal structure. Causal effect learning tasks include average causal effect estimation, heterogenous treatment effect estimation, counterfactual explanation, etc. Those tasks usually include an intervention in natural experiment, RCTs, or quasi-experimental settings. The tasks are intended to understand how the intervention affects the targeted outcome variable when all other things are equal. This is the classical ceteris paribus question in economics extended to the field of ML. For example, EconML[15] is a Python toolbox that can estimate heterogeneous

[15] https://www.microsoft.com/en-us/research/project/econml/

treatment effects from observational data using ML algorithms, such as orthogonal random forests, doubly robust learners and deep IV.

Causal structure learning task, or causal discovery task, is to examine whether a certain set of causal relationships exists among the variables (Chen et al., 2022). Our model TCDF belongs to this group. A simplified description of causal structural learning can be characterized as the following. Given a set of random variables $V = \{X_1, \cdots X_n\}$, we want to discover the causal graph that represents the causal relation of all variables. If variable X_i is modified (through techniques such as permutation), another variable X_j would change significantly when all other variables were fixed at some values, then this implies X_i is a direct cause of X_j (Chen et al., 2022; Schölkopf, 2022). The causal graph can be used to identify the effect that would occur in other variables when the value of a variable is changed (a treatment or intervention) on experimental data, it can also be used on observational data to discover a set of cause-effect relations among all the variables.

Here are some examples of causal structure learning tools. CausalNex[16] is a Python library that leverages Bayesian Networks to identify causal relations, it can encode or augment domain expertise into the graph model to conduct counterfactual analysis. CausalDiscovery[17] is another library that includes 17 algorithms for graph skeleton identification and 19 algorithms for causal directed graph prediction, including 10 graphical and 9 pairwise approaches. The goal of CausalDiscovery is "to learn the causal graph and the associated causal mechanisms from the samples of the joint probability distribution of the data" (Kalainathan and Goudet, 2019). The TCDF used in our paper is another example, which can be used on time series data to uncover pairwise cause-effect relations in observational data.

[16] https://github.com/quantumblacklabs/causalnex
[17] https://github.com/FenTechSolutions/CausalDiscoveryToolbox

Causal structure learning is an ideal ML tool for uncovering causal structures in a dataset. This data-driven approach is model-free and assumption-free in most cases. It is suitable for large and complex data, and also traditional structured data. For cross-sectional data, it can be used for counterfactual explanation. For time-series data, it can identify the temporal precedence in pairs of variables, which makes them the perfect candidate for studying instantaneous and lagged interdependences among a set of variables.

The research question of this paper is to understand the contagion and spillovers during the eurozone sovereign crises. We define contagion as the instantaneous effect from a source country (in our case, crisis country) to another country, spillovers are the lagged effect. The effects are observed through co-movements in the CDS market, characterized as pairwise cause-effect relationships. Out of all the current ML toolbox, the causal structure learning algorithms fit this question the best. It is by far the most powerful set of tools in ML that have the ability for causal discovery.

The reason that we use TCDF out of all other causal structure learning algorithms is because, at the time of writing this paper, TCDF is one of the first algorithms to incorporate deep learning into the temporal causal discovery framework. Previous models such as Bayesian causal structure learning algorithms, do not have the representation power of neural networks. TCDF uses convolutional neural networks (CNN) instead of the commonly used recurrent neural network (RNN), thus it evades the vanishing gradient problem often associated with RNN. Also, CNN can automatically detect the important features in the data through backpropagation (Yamashita et al., 2018). By interpreting the internal parameters of the CNN, TCDF can not only discover the instantaneous cause-effect relationships, but also the delayed cause-effect relationships.

Another reason to use TCDF is that it is designed for continuous time-series data, while the majority of causal structure learning algorithms only apply to i.i.d cross-sectional data (e.g.,

CausalNex, pcalg[18]). In temporal causal discovery methods, many methods cannot tolerate non-stationarity nor nonlinearity, whereas TCDF can handle those data issues. TCDF has also shown an outstanding performance on discovering causal relations in an experiment using financial time series data (Nauta et al., 2019), making it the perfect candidate to study the CDS data.

Comparing to econometric methods, TCDF is model-free and assumption-free, it can pick up complex nonlinear and dynamic relationships in the data, it can also withstand non-stationarity and heteroscedasticity (those are the features of CNN). TCDF can discover the exact time lag between the cause and its effect, and potential confounding factors in the cause-effect relationships. In a nutshell, TCDF is a flexible and powerful ML method, and it aligns perfectly with the research questions of this paper.

Figure 2.8 shows the steps of a TCDF algorithm. In order to learn a temporal causal graph from time series data, it first performs a time series prediction with N independent CNNs $N_1, \cdots N_n$, all having time series $X_1, \cdots X_n$ as input. It then uses the attention scores obtained from the prediction to run Attention Interpretation, after that, it will run the following two steps in parallel: Causal Validation and Delay Discovery.

In the first step, TCDF uses Attention-based Dilated Depthwise Separable Temporal Convolutional Networks (AD-DSTCNs) to predict the time series, the network N_j is trained to predict X_j. During this step, one needs to select the number of hidden CNN layers and the kernel width to train the networks. The AD-DSTCNs have incorporated an attention mechanism which can produce a vector of attention scores a_j for each input time series. In the second step, the attention scores are compared across all input time series to determine the inputs that are the potential causes for each input time series. Each time series now has a set of potential causes.

[18] https://cran.r-project.org/web/packages/pcalg/index.html

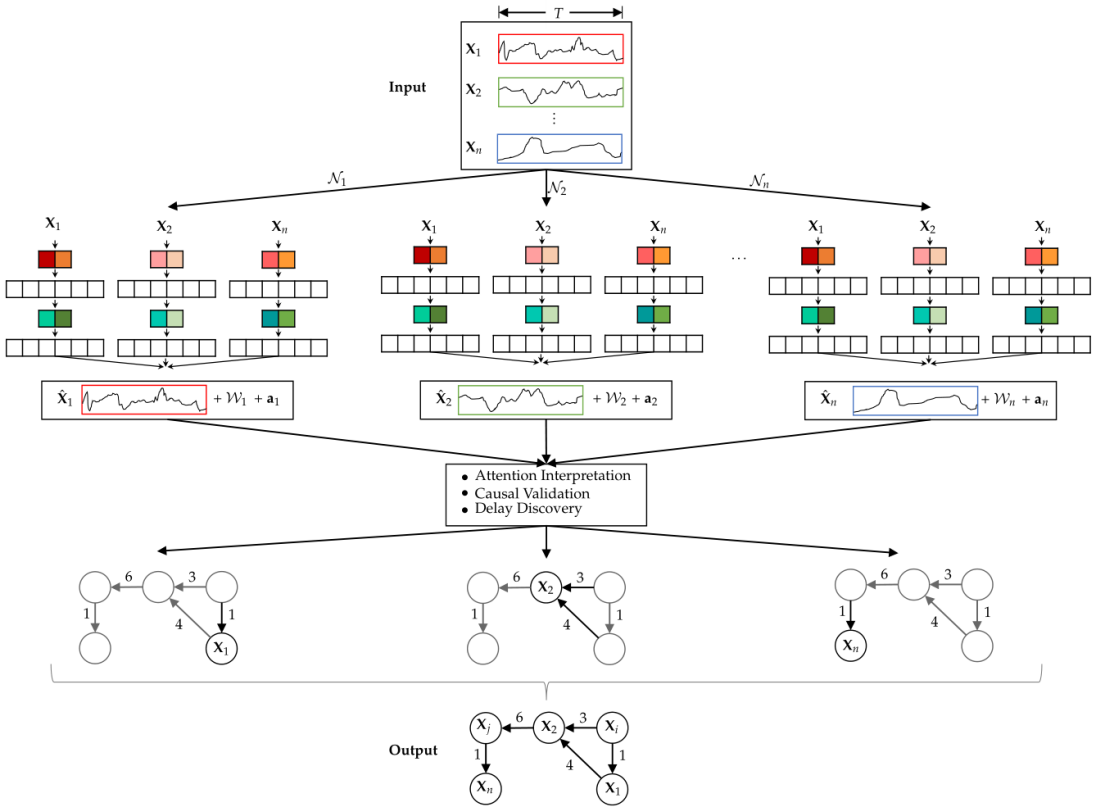

Figure 2.8. Architecture of the TCDF method, from Nauta et al. (2019).

In step three, TCDF uses the Permutation Importance Validation Method (PIVM) to run causal validation. For each target time series, PIVM creates an intervened dataset for each potential cause where the values of the potential cause are randomly permuted, then runs the trained network in step one on the intervened dataset to predict the target time series and measures the intervention loss. If loss exists, then it implies that the potential cause is the actual cause for the target time series. Parallelly, in step four, since TCDF uses a depthwise separable architecture, when a real cause is detected, the kernel weights of the AD-DSTCN for the cause time series and effect time series can be used to infer the lag in the cause-effect relationship. Finally, TCDF merges the results from these four steps to create a causal graph that shows the discovered causal relationships and their delays. TCDF can also detect the existence of hidden

confounder between two time series (see Nauta et al. (2019) for a detailed explanation of the TCDF algorithms).

TCDF can discover cause-effect relationships for both instantaneous and delayed causes in time series data. In the context of the eurozone crisis, this cause-and-effect detection is translated into contagion and spillovers among the European countries.

2.5.2 Granger Causality

In order to compare the ML methods with a benchmark econometric model, we also use pairwise Granger causality to study the contagion and spillovers among the European countries. In the eurozone crisis literature, it is common to use vector autoregression (VAR) to study the contagion and spillovers (e.g., Kalbaska and Gątkowski, 2012; Gómez-Puig and Sosvilla-Rivero). Some of them adopt a panel VAR approach, since the macroeconomics fundamental variables are included in their data. In the broader literature on contagion and spillovers, Granger causality is widely used. For example, Nagayasu (2001) uses Granger causality to study the spillovers between the foreign exchange market and the stock market in the Philippines and Thailand. Bekiros (2014) use Granger causality to study the volatility spillovers from the U.S. to the BRIC markets during the 2008-2009 financial crisis. We pick Granger causality because it is the benchmark tool in the contagion and spillovers literature.

Granger causality is a term named after Nobel laureate Clive Granger. If one time series X is useful in forecasting another time series Y, then one can say that X Granger cause Y (Granger, 1969). The Granger causality test can be conducted by regressing Y on lagged values of X and lagged values of Y, if lagged values of X can provide statistically significant information about values of Y, then X can Granger cause Y.

This paper uses a bivariate vector autoregressive (VAR) model of time series X_1 and X_2 to test the pairwise Granger causality between two countries.

$$X_{1,t} = \sum_{j=1}^{p} A_{11,j} X_{1,t-j} + \sum_{j=1}^{p} A_{12,j} X_{2,t-j} + E_{1,t} \qquad (2.1)$$

$$X_{2,t} = \sum_{j=1}^{p} A_{21,j} X_{1,t-j} + \sum_{j=1}^{p} A_{22,j} X_{2,t-j} + E_{2,t} \qquad (2.2)$$

In equation 2.1 and 2.2, X_1 and X_2 stands for the two CDS spreads time series in a country pair, p is the maximum number of lags (model order), the vectors A_{11}, A_{12}, A_{21}, A_{22} contain the coefficients of the model, E_1 and E_2 are the residuals for each time series. X_1 can Granger cause X_2 if the coefficients in A_{21} are jointly significantly different from zero, similarly, X_2 can Granger cause X_1 if the coefficients in A_{12} are jointly significantly different from zero.

After fitting the VAR model, one can implement two causality tests. The first is a F-test for pairwise Granger causality, whether the coefficient in A_{12} or A_{21} are jointly zero. The second is a Wald-type test characterized by testing for nonzero correlation between the error processes of the two time series (see Pfaff and Stigler (2021) for details).

These two tests are correspondent to the definition of spillovers and contagion in this paper. The Granger causality test can detect whether the past information of a time series is useful in forecasting another. This is commensurable to our definition of spillovers, that is the lagged transmission of a shock. The Wald-type test can detect the instantaneous correlation between the error processes, but not the direction of causality. This instantaneous causality is not a perfect measure for contagion (instantaneous transmission of a shock), but it can supplement the Granger causality results.

For VAR, it is important to choose the proper lag length. We estimated the optimal lag lengths using various selection criteria: Akaike information criterion (AIC), Schwarz information criterion (SIC), Hannan-Quinn criterion (HQC) and final prediction error (FPE). The optimal lengths are similar according those criteria, we use the AIC results in this paper.

The existing literature has different choices of the optimal lag lengths in VAR. For example, Koutmos (2018) uses weekly CDS data to test for pairwise Granger causality in bivariate VAR, the author report results of all country pairs with lag 1, 2, 3, and 4. Kalbaska and Gątkowski (2012) use a multivariate VAR that includes all countries on weekly CDS, they choose a lag of 3 for pre-crisis and 6 for the crisis period according to AIC. Different from the above papers, our paper will use the optimal lag lengths for each country pairs and report the results for phase 2 and 3. Phase 1 and phase 4 have too many Granger causal relationships so they are not reported.

If we allow unequal lag lengths of X_1 and X_2 in the VAR, that is to say, we relax the assumption that $p = q$ and $r = s$ in equations 2.3 and 2.4. This becomes the Hsiao's version of Granger causality (Hsiao, 1981).

$$X_{1,t} = \sum_{j=1}^{p} A_{11,j} X_{1,t-j} + \sum_{j=1}^{q} A_{12,j} X_{2,t-j} + E_{1,t} \qquad (2.3)$$

$$X_{2,t} = \sum_{j=1}^{r} A_{21,j} X_{1,t-j} + \sum_{j=1}^{s} A_{22,j} X_{2,t-j} + E_{2,t} \qquad (2.4)$$

Hsiao's version of Granger causality is a step-wise procedure based on Granger's concept of causality, this method can identify the optimal lag for each bivariate autoregression with different lags for the variables. By comparing the AIC or FPE of the univariate regression and bivariate regression, one can infer causality between the variables. This approach is an extension of the original Granger causality, but they differ in the testing methods. The results from the Hsiao's version of Granger causality test vary greatly from the original Granger causality test. Since the focus of this chapter is on the comparison between TCDF and Granger causality, we will only report the optimal lags between all the country pairs, but not the test results of the Hsiao's version.

2.6 Results

2.6.1 TCDF results

In the TCDF, there are two hyperparameters that need to be fine-tuned for the CNN: the number of hidden layers and the kernel width. For the number of hidden layers, generally, a hidden layer of 2 works best if one has a small dataset, increasing the layer number may cause overfitting. For the kernel width, a common suggestion for CNN is 3. After training several networks with different layer numbers and kernel widths, we pick a hidden layer of 3 and kernel size of 3 according to domain knowledge. Specifically, this paper has trained 4 networks with the same level of hidden layer and kernel size to make the results comparable.

Though there are 4 phases in the CDS dataset, in the pre-crisis period (phase 1), there is too little data variation, as shown in both Figure 2.7 and Table A.3 panel A, so phase 1 is excluded from the TCDF analysis. The first causal graph (Figure 2.9) shows the results for the eurozone countries from 1 August 2007 to 30 April 2010 (phase 2). This is the crisis buildup period. We have 12 countries for phase 2 (the U.K. is excluded due to data availability). The countries are put in circles, the lines with arrowheads between the circles show pairwise causal direction. The number on the line is the lag of the cause-effect relationship, the unit of the lag is day (our CDS data excludes weekends).

A question one might ask is why not train separate networks for different groups of countries? For example, a causal graph for only the 5 crisis countries or the 6 core countries? This is because of the black box nature of ML methods. If one performs TCDF on those subgroups of countries during the same periods, the results between the groups are not comparable to each other. For the purpose of consistency, we only report causal graphs of all countries in a given period.

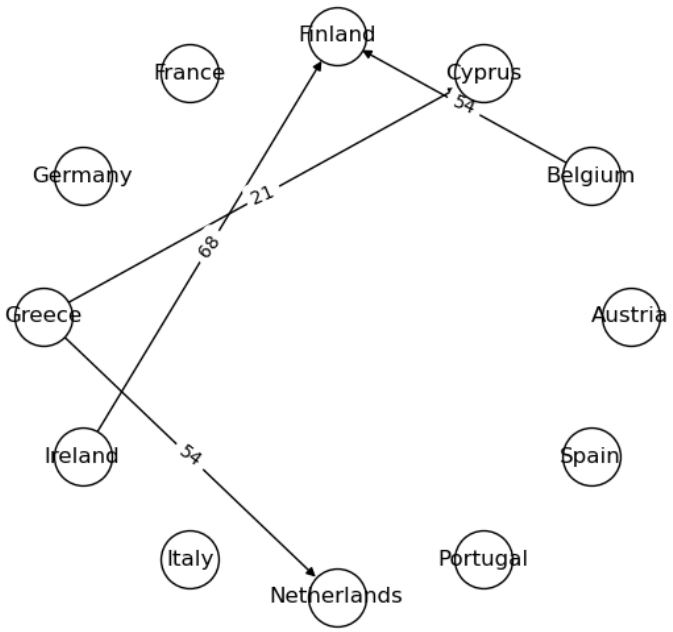

Figure 2.9. Temporal causal graph of 12 Eurozone countries from 1 August 2007 to 30 April 2010 (phase 2).

In Figure 2.9, there are four cause-effect relationships. Greece affects Cyprus and the Netherlands, with a lag of 21 and 54. Belgium and Ireland both impacts Finland, the lag length of Belgium is 54, Ireland is 68. There is no detected cause-effect relationship or confounder in the other 8 countries. Within the periphery eurozone countries, only Greece has an impact on Cyprus during the pre-crisis period.

It is not surprising that Cyprus is affected by Greece. The bank of Greece reports the net inflows of foreign direct investment into Greece during the period from 2002 to 2010 per country of origin[19]. Cyprus ranks 2nd out of all countries, the Netherlands ranks 5th. Cyprus also holds a large amount of Greek assets (mainly Greek sovereign bonds) during this period (Zenios, 2013). This may explain the lagged impact of Greek CDS on Cyprus and the Netherlands.

About the impact from Belgium and Ireland to Finland, we could not find relevant studies on this subject, a possible channel for the causal impact is through the banking sector. Ireland

[19] https://www.bankofgreece.gr/en/statistics/external-sector/direct-investment/direct-investment---flows

has the largest banking sector in the EU, Belgium's banking sector is the 3rd largest (see Figure 2.3). Such pan-European banks can insert an effect on the Finnish CDS through holding of Finnish sovereign bonds.

The second result is the causal graph (Figure 2.10) for all 13 countries from 3 May 2010 to 29 March 2013 (phase 3). Phase 3 is the eurozone crisis period; there is a lot more going on in this graph compared to phase 2. This increase in the number of cause-effect relationships during the recession period is an indication of contagion during the crisis period.

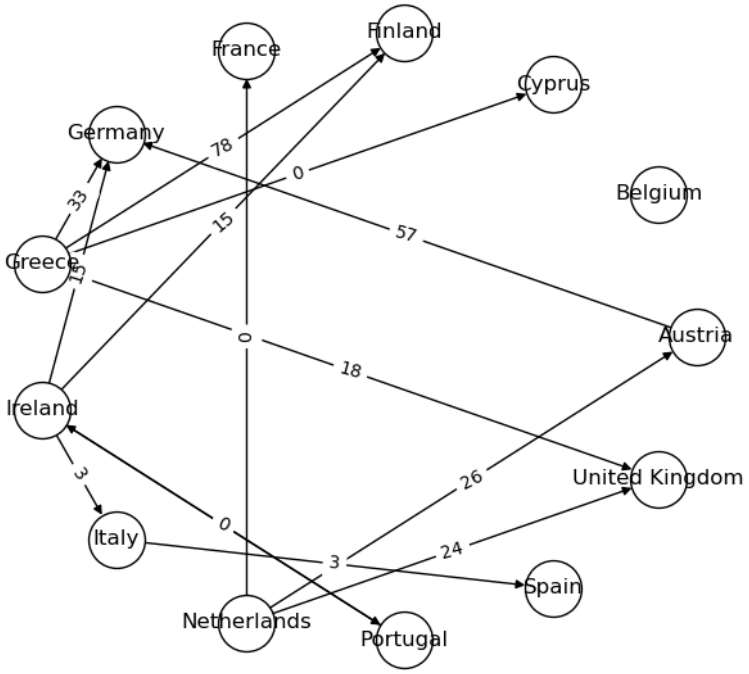

Figure 2.10. Temporal causal graph of all 13 countries from 3 May 2010 to 29 March 2013 (phase 3).

In Figure 2.10, there are 11 cause-effect relationships and 1 hidden confounder, 7 out of the 11 relationships involve a periphery eurozone country. The double arrow-headed line between Ireland and Portugal suggests that the co-movement in their CDS spreads is caused by a hidden confounder, the confounder is not any other countries' sovereign CDS spread. Greece, who is

the center of this graph and the center of the crisis, has impacts on four other countries: Cyprus, Finland, Germany, and the U.K., the lags are 0, 78, 33, 18, respectively. The Greece-Cyprus relationship carries on from phase 2 to phase 3, but the lag decreases from 21 to 0, showing a change from delayed spillovers to an instantaneous contagion from Greece to Cyprus, this is likely due to the holding of Greek sovereign bonds in the Cypriot banks (Zenios, 2013). During the crisis periods, the shocks of the Greek sovereign bonds speed up the transmission, shorten the lag between Greece and Cyprus. Though Greece exhibits its influence in Europe during phase 3, it does not affect the other crisis countries besides Cyprus.

Ireland affects Finland, Germany and Italy, with lags of 15, 15 and 3. Ireland's failing banking sector has an impact on many other eurozone countries because of its size. Ireland has an almost instantaneous influence on Italy. Surprisingly, given the geographical proximity and economic integration between Ireland and the U.K., Ireland does not affect the U.K. during the crisis. This result shows that countries within the eurozone are more vulnerable to shocks from the member states.

Italy, who is affected by Ireland, has an influence on Spain with a lag of 3. This is an interesting chain of cause and effect that Ireland affects Italy with a lag of 3, then Italy affects Spain with a lag of 3. The impact from Italy to Spain can also be found in other phases. The existing literature presents numerous evidences of the interdependency between Italy and Spain (e.g., Broto and Pérez-Quirós, 2015). In our CDS data, we can only observe the spillovers from Italy to Spain, but not the other direction.

The Netherlands, who has solid fiscal fundamentals during phase 3, affects France instantaneously. It also affects Austria and the U.K. with lags of 26 and 24. The last piece of the causal graph is Austria, who affects Germany with a lag of 57. Those results are new findings to the existing literature, but it is challenging to understand the interconnection among those countries by using the CDS data alone.

Out of the 13 countries, Greece has the largest influence, Germany receives the most impact. Being the largest exporter inside the European Union, Germany is affected by Greece and Ireland. One might speculate that the channel of transmission could also be through the export market. Both Greece and Ireland affect Finland. This confirms the transmission from the periphery countries to the core countries inside the euro area. Though this paper only includes one country outside of the eurozone, the U.K. receives influence from Greece as well. This insinuates the transmission from the eurozone countries to the European Union countries outside of the eurozone.

All the 6 periphery eurozone countries are in one or more cause-effect relationships. This is certainly the sign of contagion and spillovers among the periphery eurozone countries. In the eurozone crisis literature, an important question is whether Greece is the origin of other countries' sovereign crises. Our results show that Greece only has an influence on Cyprus, but not on any other periphery eurozone country. This finding concurs with the crisis consensus in section 2.2 of this chapter. The nature of each country's crisis is different. The only country that has a pure sovereign crisis is Greece. The Cypriot banking crisis and sovereign crisis have their roots in holding too many Greek assets, this explains the instantaneous cause-effect relationship between the two countries during phase 3. Ireland and Spain's problems root in their banking sectors, which are not directly linked to the falling Greek sovereign. Portugal and Italy have long-term slow economic growth and loss of competitiveness, they have a special "balance of payment crisis in a monetary union", which is almost independent from what happens in Greece.

The third result is the causal graph (Figure 2.11) of all 12 eurozone countries from 1 August 2007 to 29 March 2013 (phase 2 and phase 3). This period encompasses two major crises: the subprime crisis and the eurozone crisis. Therefore, the cause-effect relationships detected in such a long period do not reflect a transmission during a crisis, but rather the close economic

interdependences between the countries. Hence, Figure 2.11 can serve as a reference to other causal graphs.

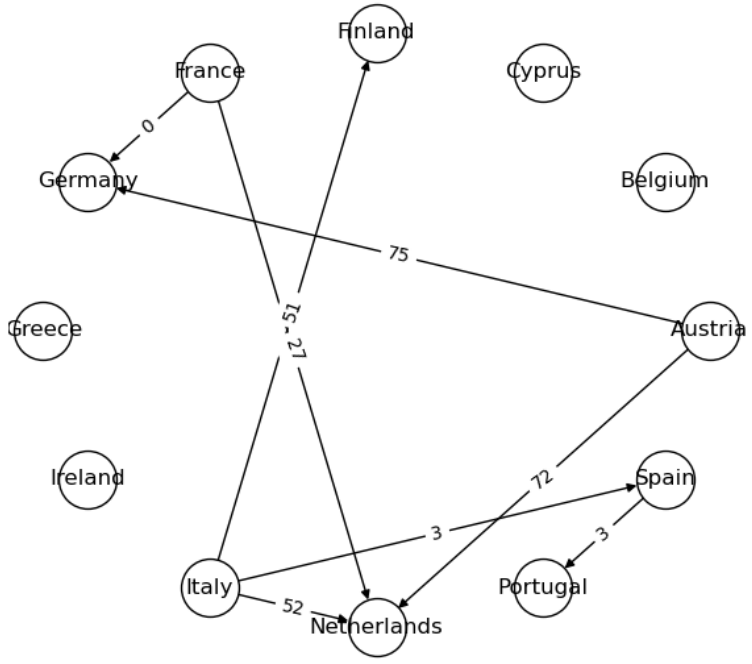

Figure 2.11. Temporal causal graph of 12 Eurozone countries from 1 August 2007 to 29 March 2013 (phase 2 and phase 3).

Germany and France are the largest and second largest country in the eurozone. As the center of the European Union, these two countries join action in many areas, which explains the instantaneous relationship from France to Germany in Figure 2.11. Another chain of causal relationship is from Italy to Spain, then to Portugal, with a lag of 3 in both. This is an indication of regional integration of the southern peripheral economies. The cause-effect relationship between Italy and Spain is the same as Figure 2.10, that Italy affects Spain with a lag of 3, this relationship carries on to phase 4 as well. The other cause-effect relationships all have larger lags, which makes them less important in this context.

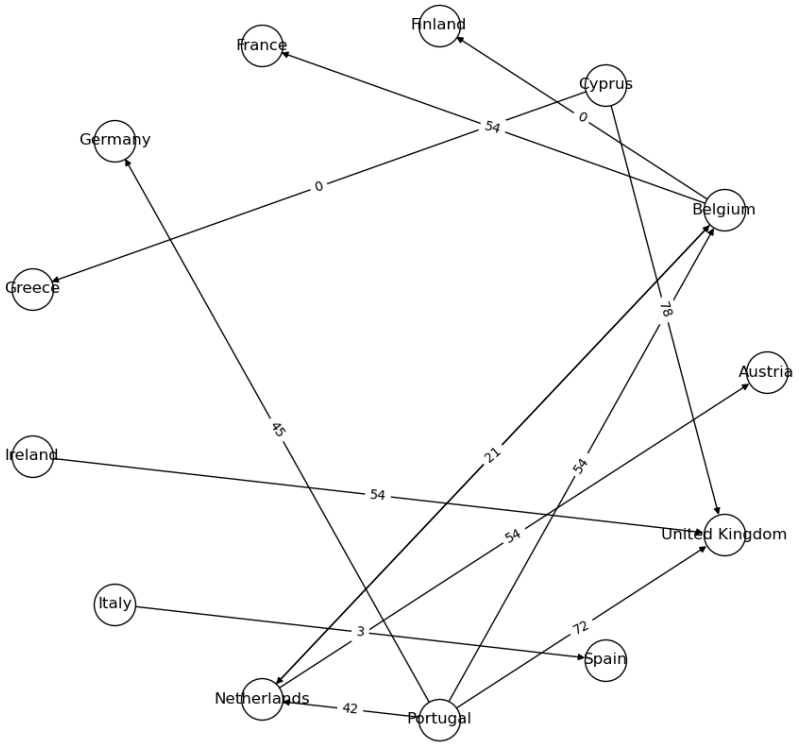

Figure 2.12. Temporal causal graph of all 13 countries from 1 April 2013 to 31 December 2015 (phase 4).

The fourth result is for the post-crisis recovery period after the eurozone crisis. Figure 2.12 shows the causal graph of all 13 countries from 1 April 2013 to 31 December 2015 (phase 4). Interestingly, there are 11 cause-effect relationships and 1 hidden confounder, just like Figure 2.10. Belgium and the Netherlands affect each other with a lag of 21. This suggests the existence of a hidden confounder for those two countries during the recovery period. In the post-crisis periods, there is another "self-fulfilling" story of recovery. In this round, the previously distressed crisis economies are injecting vigor into Europe.

In Figure 2.12, Portugal is the most influential country in phase 4, it affects 4 non-crisis countries: Belgium, Germany, the Netherlands, and the U.K., with lags of 54, 45, 43 and 72.

Portugal indeed has a remarkable post-crisis recovering rate (Reis, 2015), the TCDF results show that the bourgeoning Portuguese economy has spilled its vitality into the core European countries with a lag of roughly 2 months.

Cyprus has an impact on Greece and the U.K., with lags of 0 and 78. The change of direction between Cyprus and Greece is of particular interest. In phase 2 and phase 3, Greece affects Cyprus during the economic downturns, but Cyprus affects Greece during the economic recovery in phase 4. The lag between the two is 0 in both phase 3 and phase 4, their close link through the Greek assets in the Cypriot banks remains the same.

In phase 4, Ireland affects the U.K. with a lag of 54, this is new information since Ireland does not affect the U.K. in phase 2 and 3. Italy affects Spain with a lag of 3 (same result as in Figure 2.10 and Figure 2.11). There are other cause-effect relationships among the core eurozone countries, but they are unrelated to the topic of contagion and spillovers.

2.6.2 Granger causality results

This paper uses bivariate VAR to detect the Granger causality using a F-test, and a Wald-type test for instantaneous causality in country pairs. For each VAR, we use AIC, SIC, HQC and FPE for optimal lag selection, all the criteria give similar results. We choose the AIC results to conduct the F-test and the Wald test.

Table 2.7 shows the optimal lag length for pairwise Granger causality using AIC during the crisis period (phase 3), the optimal lag length for different country pairs varies from 1 to 8. The lags between Greece and other countries are 1 or 2. The lags among the periphery eurozone countries are almost 1, 2 and 3, the only exception is between Ireland and Spain, which is 8.

Table 2.7. Optimal lag length for pairwise Granger causality using AIC in phase 3.

	AT	BE	CY	FI	FR	DE	EL	IR	IT	NL	PT	ES	UK
Austria	..	1	2	1	1	1	1	3	3	1	2	5	1
Belgium	1	..	3	2	3	2	1	3	4	1	2	4	3
Cyprus	2	3	..	2	3	3	2	2	2	2	2	3	2
Finland	1	2	2	..	1	1	1	3	3	2	2	3	1
France	1	3	3	1	..	1	1	3	3	3	3	7	3
Germany	1	2	3	1	1	..	1	6	4	2	2	6	1
Greece	1	1	2	1	1	1	..	1	2	1	1	2	1
Ireland	3	3	2	3	3	6	1	..	3	3	6	8	3
Italy	3	4	2	3	3	4	2	3	..	2	2	2	3
Netherlands	1	1	2	2	3	2	1	3	2	..	2	2	2
Portugal	2	2	2	2	3	2	1	6	2	2	..	2	2
Spain	5	4	3	3	7	6	2	8	2	2	2	..	3
The U.K.	1	3	2	1	3	1	1	3	3	2	2	3	..

Please see Table A.1 for the country codes in the appendix.

Table 2.8 and Table 2.9 report the test results of phase 2 (12 eurozone countries) and phase 3 (all 13 countries). In phase 1 and phase 4, the results show that almost all countries can Granger cause all other countries. For simplicity, the results of phase 1 and phase 4 are not reported. The top row of the tables are the cause countries, the numbers in red indicate an at least 5% significance level that the cause country can Granger cause the countries listed in the first column. For instance, in column 2 of Table 2.8, Austria can Granger cause all other countries, in column 4, Cyprus can only Granger cause Finland. This result echoes with the correlation matrix in Table A.3. Cyprus, being the smallest economy of all, lacks the strengths to affect bigger countries. In phase 2, almost all countries (except Cyprus and Finland) can Granger cause all other countries.

In the crisis period (Table 2.9), one of the important research questions is whether Greece is the origin of other countries' sovereign crises. Our results show that Greece does not Granger cause any other country, including the crisis country and the non-crisis country. Italy and Spain

Table 2.8. Granger causality test for the crisis buildup period (phase 2)

	AU	BE	CY	FI	FR	DE	EL	IR	IT	NL	PT	ES
Austria	NA	0.00	0.98	0.16	0.01	0.01	0.01	0.00	0.03	0.29	0.09	0.00
Belgium	0.00	NA	0.56	0.12	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
Cyprus	0.00	0.00	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Finland	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
France	0.00	0.00	0.40	0.15	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Germany	0.00	0.00	0.41	0.03	0.00	NA	0.00	0.00	0.00	0.02	0.00	0.00
Greece	0.00	0.00	0.09	0.08	0.00	0.00	NA	0.00	0.01	0.03	0.00	0.00
Ireland	0.00	0.00	0.58	0.01	0.00	0.00	0.00	NA	0.00	0.01	0.00	0.00
Italy	0.00	0.00	0.12	0.81	0.00	0.00	0.00	0.00	NA	0.06	0.01	0.00
Netherlands	0.00	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00	NA	0.00	0.00
Portugal	0.00	0.00	0.36	0.01	0.00	0.00	0.00	0.00	0.00	0.00	NA	0.00
Spain	0.00	0.00	0.73	0.05	0.00	0.00	0.00	0.00	0.01	0.04	0.01	NA

Red indicates significance at 5% level.

Table 2.9. Granger causality test for the crisis period (phase 3)

	AU	BE	CY	FI	FR	DE	EL	IR	IT	NL	PT	ES	UK
Austria	NA	0.00	0.01	0.05	0.00	0.00	0.23	0.04	0.00	0.00	0.31	0.00	0.01
Belgium	0.68	NA	0.02	0.01	0.00	0.11	0.16	0.97	0.02	0.15	0.66	0.00	0.33
Cyprus	0.00	0.00	NA	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
Finland	0.01	0.00	0.04	NA	0.00	0.00	0.05	0.22	0.00	0.00	0.02	0.00	0.00
France	0.13	0.13	0.03	0.55	NA	0.31	0.01	0.24	0.03	0.22	0.26	0.00	0.47
Germany	0.30	0.00	0.05	0.32	0.01	NA	0.09	0.91	0.02	0.02	0.02	0.08	0.17
Greece	0.00	0.00	0.25	0.00	0.00	0.01	NA	0.00	0.02	0.00	0.00	0.00	0.01
Ireland	0.01	0.10	0.15	0.21	0.00	0.01	0.97	NA	0.00	0.02	0.03	0.00	0.07
Italy	0.03	0.06	0.12	0.06	0.01	0.00	0.15	0.62	NA	0.85	0.17	0.04	0.09
Netherlands	0.52	0.00	0.06	0.00	0.00	0.00	0.10	0.05	0.00	NA	0.10	0.00	0.24
Portugal	0.02	0.00	0.13	0.39	0.01	0.07	0.48	0.30	0.00	0.04	NA	0.00	0.01
Spain	0.01	0.37	0.60	0.30	0.02	0.13	0.07	0.79	0.00	0.20	0.67	NA	0.00
The U.K.	0.83	0.02	0.00	0.02	0.00	0.00	0.13	0.06	0.00	0.17	0.04	0.00	NA

Red indicates significance at 5% level.

can Granger cause all other countries including the U.K. This implies that there are spillovers within the periphery eurozone countries, from the periphery eurozone countries to the core eurozone countries, and to outside of the eurozone. The Granger causality results of the periphery eurozone countries are summarized in Table 2.10.

Table 2.10. Granger causality results for the periphery eurozone countries in phase 3

Cause Country	**can Granger cause:**
Cyprus	None
Greece	None
Ireland	Cyprus, Greece
Italy	Cyprus, Greece, Ireland, Portugal, Spain
Portugal	Cyprus, Greece, Ireland
Spain	Cyprus, Greece, Ireland, Italy, Portugal

From the above tables, we notice that a lot of Granger causality has been detected compared to the TCDF results for the lagged spillovers. Though Granger Causality is not equivalent to causality, it is closely linked with causality in settings such as a VAR (White et al., 2011). It can provide potential points of interest for further investigation; some spurious relationships might be included as well (He and Maekawa, 2001). A multivariate panel VAR with macroeconomic variables would improve upon these preliminary results.

To compare the TCDF and the Granger causality regarding contagion, we also report some instantaneous causality results from the VAR. The instantaneous causality is an indicator of immediate transmission of a shock during the crisis. Table 2.11 shows all the instantaneous cause-effect relationships detected in the TCDF, and reports the Wald test results for instantaneous causality. 2 out of the 7 instantaneous cause-effect relationships in the TCDF

cannot be found using the VAR. During the crisis period (phase 3), Greece has no instantaneous causality towards Cyprus.

Table 2.11. Instantaneous causality Wald tests

Phase 3	EL cause CY	IR cause PT	PT cause IR	NL cause FR
	N	*Y*	*Y*	*Y*
Phase 2+3	FR cause DE			
	Y			
Phase 4	CY cause DE	BE cause FI		
	N	*Y*		

Table 2.12. Optimal lag for Hsiao's version of Granger causality for the crisis period (phase 3)

	AU	BE	CY	FI	FR	DE	EL	IR	IT	NL	PT	ES	UK
Austria	..	1, 1	5, 2	5, 2	1, 1	1, 1	1, 1	1, 1	1, 1	1, 1	1, 1	1, 1	1, 1
Belgium	1, 1	..	3, 5	1, 5	1, 1	1, 1	1, 1	1, 1	2, 6	1, 1	1, 1	2, 6	1, 1
Cyprus	1, 5	1, 5	..	1, 5	1, 5	1, 5	2, 5	1, 5	1, 5	1, 5	1, 5	3, 2	1, 5
Finland	1, 4	1, 5	1, 1	..	1, 1	1, 1	1, 1	2, 5	5, 2	2, 5	5, 2	1, 3	3, 5
France	3, 1	1, 1	3, 5	1, 5	..	1, 1	3, 1	1, 1	1, 1	1, 1	5, 2	1, 1	4, 2
Germany	2, 2	2, 2	5, 1	5, 1	1, 1	..	1, 1	2, 2	4, 2	2, 1	2, 2	4, 1	2, 1
Greece	1, 5	1, 5	1, 5	1, 5	1, 5	1, 5	..	1, 5	1, 5	1, 5	1, 1	1, 5	1, 5
Ireland	5, 3	3, 5	1, 6	5, 4	5, 2	10,4	5, 1	..	5, 2	3, 5	3, 2	5, 3	3, 5
Italy	1, 5	3, 6	5, 1	1, 5	5, 1	3, 5	2, 5	3, 3	..	1, 5	1, 3	2, 5	1, 5
Netherlands	1, 1	1, 1	3, 5	5, 2	1, 1	1, 1	1, 1	1, 2	1, 1	..	1, 1	1, 1	1, 1
Portugal	1, 9	10,4	1,10	10,5	10,2	10,4	10,1	9, 2	10,2	2,10	..	5, 2	2, 9
Spain	4, 5	8, 2	5, 1	2, 7	5, 1	7, 4	5, 2	1, 4	1, 4	7, 1	3, 2	..	5, 3
The U.K.	1, 1	1, 1	7, 1	7, 3	7, 2	7, 1	7, 1	7, 2	1, 1	1, 7	2, 4	2, 4	..

For Hsiao's version of Granger causality, the optimal lags of country pairs in the crisis period are reported in Table 2.12. The top row of the panels are the cause countries, the columns are the countries that receive the impact. For example, column 2 shows the lag lengths when Austria is the cause country, between Austria and Belgium, both lag length is 1, between Austria and Cyprus, the lag for Austria is 1, while the lag of Cyprus is 5. From Table 2.7 and Table 2.12,

we can see that the optimal lag lengths for country pairs vary greatly for Granger causality and Hsiao's version of Granger causality. The results of the Hsiao' tests show no causality between most country pairs. Since the results of Hsiao's test and Granger causality test are quite different, we will focus on Granger causality and not report the Hsiao's results.

2.6.3 Comparison

There are four main differences between the TCDF results and Granger causality results. First, we can detect a lot more Granger causality pairs than the TCDF, especially in phase 2 and phase 4. About 90% of the country pairs show a Granger causality in phase 2, but there are only 4 cause-effect relationships in the TCDF results. An explanation for the large amount of causality is that the bivariate Granger causality test can be affected by an omitted variable, thus showing many spurious relationships (He and Maekawa, 2001). Second, in each phase, the Granger causality results differ greatly from the TCDF results. For example, during the crisis period, there is no Granger causality nor instantaneous causality from Greece to any crisis country. But in the TCDF results, Greece can affect Cyprus instantaneously. Also, Ireland can Granger cause Cyprus and Greece, but the TCDF results show that Ireland has an instantaneous effect on Portugal and a lagged effect on Italy.

Third, the lags of the causality are different. The TCDF can detect long lags (e.g., 51, 78) between country pairs where the Granger causality optimal lag lengths are mostly between 1-10. Fourth, in all lagged cause-effect relationships found in the TCDF, if we use the lag length of the TCDF to conduct a Granger causality test in those country pairs, we can only detect some Granger causality, but not all.

In phase 2, all TCDF pairwise cause-effect relationships are significant in Granger causality results. In phase 3, 5 out of the 12 cause-effect relationships are not detected by Granger causality. In phase 2 and 3, 2 out of the 8 cause-effect relationships are not detected by Granger

causality. In phase 4, 7 out of the 11 cause-effect relationships are not detected by Granger causality.

To sum up, the Granger causality test can pick up a larger amount of causality among the countries, but the larger amount only laps over part of the cause-effect relationships picked up by the TCDF. The TCDF results are more consistent with the domain knowledge.

Many factors can contribute to the differences in the test results. For ML algorithms, they are well-known for their ability to pick the nonlinear patterns in the data, while the VAR can only pick up linear relationships. The TCDF algorithm uses the time series of all countries to fit the convolutional neural networks for the prediction task, while the pairwise Granger causality test only uses the time series of two countries. Because of the depthwise separable architecture of the TCDF, it can detect the lag between a cause and an effect, it can detect both the instantaneous (lag = 0) and delayed (lag > 0) causal relationships, while the Granger causality test leans to the lagged impact.

The lag length selection of VAR is not automated, whereas the TCDF can calculate lag lengths from 0 up to the length of the input time series automatically. The TCDF can also pick up potential confounder between country pairs. For instance, there is a confounder between Ireland and Portugal during the crisis period, a confounder between Belgium and the Netherlands with a lag of 21 after the crisis. The causal validation step in the TCDF can filter out spurious relationships in country pairs, therefore, the results of TCDF are much more succinct and comprehensive than the Granger causality approach.

2.7 Discussion

This chapter aims to study the contagion during the eurozone crisis using a novel ML method. It answers the four critical questions raised in the introduction of this chapter. The first question is “are there contagion and spillovers during the eurozone crisis?” The answer is yes. There is a

sharp increase in the number of cause-effect relationships from the crisis buildup period (phase 2) to the actual crisis period (phase 3). According to the first group's definition of contagion in section 2.1, the increase itself marks the existence of contagion during the eurozone crisis. Also, in Figure 2.10, the 11 pairs of cause-effect relationships and 1 confounding factor exhibit the close links among the European economies, for instance, there is an instantaneous effect from Greece to Cyprus, and delayed effect from Ireland to Italy, then to Spain.

The second question is whether Greece is the origin of all the crises in the eurozone and how the crisis countries affect each other. The TCDF results tell that Greece is not the black sheep of all the eurozone crises. Indeed, Greece has a great impact on Cyprus in the crisis buildup and actual crisis period, causing the banking crisis and sovereign crisis in Cyprus. The impact from Greece to Cyprus could also be observed from the decreased number of lags from the crisis buildup period to the actual crisis period. During the crisis period from 3 May 2010 to 29 March 2013 (phase 3), the Greek CDS spread can affect the Cypriot CDS spread instantaneously, showing a contagion between the two. Other than Cyprus, Greece does not impact any other periphery eurozone countries during the buildup and crisis period. This result confirms the finding in the eurozone crisis consensus that the sovereign crises in Ireland, Portugal and Spain are not of a fiscal nature.

During the crisis period, contagion and spillovers exist among the 6 periphery eurozone countries and among the 5 crisis countries. There is a two-way causal relationship between Ireland and Portugal, the two countries affect each other instantaneously, suggesting the existence of a confounding factor. Ireland also affects Italy with a lag of 3, Italy then affects Spain with a lag of 3. Ireland was the second country to receive a bailout program in the eurozone, the market risk of the Irish bond was transmitted to Italy and then to Spain.

The third question is whether there are contagion and spillovers from the crisis countries to the non-crisis country inside the eurozone. The answer is yes. In Figure 2.10, Greece shows its impact on Germany and Finland. Ireland has an influence on Finland. These findings confirm

the spillovers from the crisis countries to the core countries. Contagion and spillovers also exist among the core countries in the eurozone. The Netherlands, who has a high private debt like Portugal, also contributes to the risk spillovers in the CDS market. It affects France instantaneously and affects Austria and the U.K. with time lags.

The fourth question is whether there exist contagion and spillovers from the eurozone countries to non-eurozone countries in the European Union. Since the data used in this paper only includes one non-eurozone country (the U.K.), the TCDF results can only provide a partial picture of this question. Figure 2.10 shows that the U.K. is affected by Greece and the Netherlands during the crisis period, the spillovers from Greece confirm the transmission of market risk from the crisis countries to countries outside of the eurozone.

The answers to the four questions using TCDF are different with most previous studies, most of the existing literature suggest that there are no contagion and spillovers during the eurozone crisis, or that contagion and spillovers exist but the shock from Greece affects all other crisis countries. Those findings are in contrast with the TCDF results.

There are two papers that have similar results to our paper (Koutmos, 2018; Bampinas et al., 2020). These two papers find evidence that Greece can affect some periphery countries but not all crisis countries, and there are contagion and spillovers from the periphery countries to the core countries. Koutmos (2018) uses the weekly CDS spreads to model the Granger Causality between the CDS time series data using a VAR. They find that there exists lots of pairwise Granger causalities among all the European countries. The author then uses the proportional percentage of each country to represent its power to transmit risk, the percentage of a country is defined as the ratio between the number of countries that one can Granger cause over the total number of countries. Their results show that Greece affects 63% of other countries during the crisis period when the lag length of VAR is 1, this percentage is on a par with France, but lower than that of Belgium and Portugal.

Bampinas et al. (2020) uses the daily sovereign bond spreads and CDS spreads to study the cross-border and intra-market linkage in the eurozone countries from 2006 to 2018, they use bootstrap test for local Gaussian correlation to determine the existence of contagion between the markets. Their findings show that contagion occurs from the periphery eurozone countries' CDS spreads to their own bond spreads and to Belgium's bond market; the shocks in Italian and Spanish CDS spreads spill towards all other European CDS spreads.

Our TCDF results have consistent answers to the four questions with the above two papers, but the TCDF results include a lot more information that was not seen in previous literature. Mainly, there are four aspects. First, the TCDF can detect many cause-effect relationships that are not seen in previous works, for instance, we find that during the crisis period, there is a chain of causal effect from Ireland to Italy with a lag of 3, then from Italy to Spain with a lag of 3; the traditional econometric tools are not suitable for such causal discovery task. Second, we can observe the change of direction between some country pairs. The cause-effect relationship between Greece and Cyprus is found in all the causal graphs. Before and during the crisis, the shock of the Greek sovereign impacts Cyprus, but after the crisis, the impact is reversed that Cyprus affects Greece with its economic recovery.

Third, the lags of many cause-effect relationships are new to the literature. For example, Austria affects Germany with a lag of 57 (roughly two months) during the eurozone crisis, the Netherlands affects France instantaneously. The lags are less interesting topics in the eurozone crisis literature, we cannot find many relevant studies on the lags of the non-crisis countries. The TCDF results can also find the changes in the lags in different phases. For instance, Greece affects Cyprus with a lag of 21 before the crisis, then the lag becomes 0 during the crisis. Between Ireland and Finland, the lag changes from 68 to 15 before and during the crisis. Some previous studies have similar results showing intensified spillovers during the crisis period (e.g., Bekaert et al., 2005), but the results are mostly for the crisis countries. Fourth, the TCDF can also detect confounding factors between country pairs. The omitted variables can lead to bias in

the estimations in traditional econometric methods and there is usually no test for potential confounding factors. The TCDF results show some confounding factors, such as a confounding factor between Ireland and Portugal during the crisis period.

To sum up, the novel Temporal Causal Discovery Framework can detect instantaneous and delayed causal relationships in country pairs through a causality validation step. It offers a filtered yet richer understanding of the time series data. A lot of the cause-effect relationships detected by the TCDF are not seen in previous works. The TCDF results offer detailed causal graphs of the direction and lag of contagion and spillovers in the eurozone crisis. In a nutshell, the TCDF results add new findings to the existing literature, and these findings confirm the consensus view of the eurozone crisis.

2.8 Conclusion

In the existing eurozone crisis literature, there is no consensus on whether contagion exists during the 2010-2013 crisis. Though it has been a decade since the onset of the eurozone crisis, this question is still of importance in understanding the crisis dynamics inside the eurozone to prevent future crisis. The results from TCDF present evidence of contagion and spillovers during the eurozone crisis. It also shows the lag of spillovers from a cause country to the effect country. The existence of confounder between two countries can also be detected by identifying two-way cause and effect relationships.

This paper contributes to the literature in three ways. First, to the author's knowledge, this is one of the first attempts that use ML methods to study the contagion and spillovers during the eurozone crisis. Second, this paper goes beyond the common prediction problem associated with ML, we use deep learning causal graphs in the field of macroeconomics. Third, by using the novel deep learning framework, this paper provides a granular report of the eurozone crisis

contagion and spillovers, adding new findings to the repository, this is of great importance to future crisis management and macro-prudential regulations in the European Union.

The application of causal learning methods in the field of macroeconomics is still rare, as an early attempt, this paper has its limitations. Improvement can be made in the following directions. The first direction is data. Because of the lack of availability of high frequency macroeconomic variables, our paper only uses the daily CDS spreads data for causal discovery. The field of macroeconomics has seen more and more high frequency granular dataset, for example, researcher in MIT and Harvard start the Billion prices project[20] which can provide daily consumer price and monthly inflation rate in major countries around the world. In the eurozone crisis context, future research could improve by adding more macroeconomic covariates to the TCDF framework. Also, there are many cause-effect relationships that are not presented before, for example, during the eurozone crisis, the Netherlands affects France instantaneously, affects Austria and the U.K. with lags of 26 and 24. Those are interesting data patterns that are picked up by the TCDF, our CDS data cannot support a deep-dive analysis on these findings. Future research can focus on those less visited topics by utilizing more country specific variables.

Besides the macroeconomic variables, there are some high frequency data available, such as sovereign bond yield, stock market index. Such data can all be combined with the CDS data to expand the current time series observations into higher dimensional data. Moreover, our paper only studies the effects between two countries, but not among country groups. A further extension is to use country groups as features to analyze the contagion and spillovers from the crisis countries to other country groups (e.g., OECD countries, developing countries).

One drawback of the TCDF framework is that the convolutional neural nets cannot identify the exact timing of the cause and effect, one can only observe the lags between the cause-effect pairs. This is the reason why we split the entire sample into four phases to study the crisis

[20] http://www.thebillionpricesproject.com/

dynamic. Other studies that do not have a temporal measure of the spillovers also follow this approach of sample splitting. Different splitting strategies could give different results, for instance, our results for phase 2, phase 3 and phase 2+3 display similar but distinct cause-effect relationships, we find a smaller number of cause-effect relationships in the longer period (phase 2+3). The current causal discovery learning tools such as TCDF, CausalNex and pcalg do not have this capability to measure the temporal effect of transmissions, future work can improve the current models by incorporating a temporal measurement.

The TCDF belongs to causal structure learning tools, not the causal effect estimation learning tools. One can also apply the causal effect estimation learning on the CDS data. It should require a separate set of time series variable on countries outside of the EU as control group. This allows the discovery of heterogeneous effect of crisis for each country, or the crisis country as a whole. Besides measuring the magnitude of the crisis effect, the temporal dimension of the effects can also be obtained, i.e., how long does the impact last for each country? For instance, the Bayesian structural time series (bsts) model and CausalImpact can construct the counterfactual for each crisis country, then a Markov Chain Monte Carlo algorithm can be used for posterior inference to report the pointwise 95% predictive intervals of the crisis effect, the time series of pointwise intervals can provide further information of the temporal evolution of the crisis.

Other than ML methods, traditional econometric tools should be explored on this subject as well, a comparison between non-parametric estimation and the ML estimation of the spillovers can be an interesting topic. Different ML models and econometrics models can be crossed examined to provide a panorama view on this subject. Our paper provides an overview of the eurozone crisis using ML methods, and there are many other topics in macroeconomics that are dominated by the model-driven research paradigm. We believe that with the rapid development of causal learning, more advanced ML techniques could find their place in the field of macroeconomics.

Appendix A

Table A.1 Timeline of major events in the European Union.

Feb 1992	Maastricht treaty signed
Sep 1992	European exchange-rate mechanism (ERM) crisis
Oct 1993	Maastricht treaty ratified
Jun 1997	Stability and Growth Pact signed
Jan 1999	European Monetary Union (EMU) begins with 11 countries
Jan 2001	Greece joins EMU
Jan 2002	Euro notes and coins introduced
Nov 2003	Germany and France breach stability pact
Aug 2007	ECB liquidity injection begins
Jan 2008	Cyprus and Malta join the euro
Sep 2008	Lehman Brothers collapses
Jan 2009	Greece downgraded
Nov 2009	New Greek government admits to a bigger budget deficit
May 2010	First Greek bailout
June 2010	The temporary European Financial Stability Facility (EFSF) is created
Oct 2010	Deauville deal on private-sector involvement
Nov 2010	Irish bailout
May 2011	Portuguese bailout
July 2011	Second Greek bailout
Aug 2011	ECB buys Italian and Spanish bonds
Oct 2011	Haircut on Greek debt
Nov 2011	Mario Draghi becomes ECB president
Dec 2011	ECB launches LTRO. Fiscal compact treaty agreed
Feb 2012	New Spanish government admits higher budget deficit
Jun 2012	Partial bailout for Spanish banks
July 2012	Draghi gives “whatever it takes” speech
Aug 2012	ECB agrees Outright Monetary Transactions (OMT) program

Sep 2012	The European Stability Mechanism is created
Nov 2012	Greek debt burden spread out; interest rate cut
Feb 2013	Indecisive Italian election
Mar 2013	First Cyprus bailout, banks shut
May 2013	Second Cyprus bailout
Dec 2013	Ireland exits bailout program
Jan 2014	Latvia joins euro, Spain exits bailout program
June 2014	Portugal exits bailout program
Nov 2014	ECB takes over supervision of the most important banks in the eurozone
Aug 2015	Third Greek bailout
Mar 2016	Cyprus exits bailout program
Aug 2018	Greek exits bailout program

Timeline of major events in the European Union from 1992 to 2018. Adapted from Peet and La Guardia (2012).

Table A.2. Descriptive statistics of CDS spreads in basis points.

Country	Mean	Median	Std. Dev.	Min	Max
Austria	2.06	2.00304	.367	1.493	2.698
Belgium	2.45	2.49106	.277	1.926	3.119
Cyprus	7.694	7.8675	1.158	5.485	9.718
Finland	1.792	1.625	.585	1.083	3.088
France	2.02	1.8305	.445	1.47	2.81
Germany	2.028	1.84057	.52	1.292	3.204
Greece	10.29	11.3263	3.6	4.717	15.357
Ireland	2.368	2.39323	.297	1.667	3.137
Italy	9.273	9.42577	2.156	5.291	13.514
Netherlands	1.794	1.81833	.337	1.129	2.575
Portugal	6.221	6.17095	1.575	3.864	8.99
Spain	2.941	2.8709	.42	2.348	5.566

Panel A. Phase 1, from 3 October 2005 to 31 July 2007, observations = 417

Country	Mean	Median	Std. Dev.	Min	Max
Austria	59.559	58.37813	56.94	1.92	268.879
Belgium	43.593	35.61889	32.985	2.6	155.526
Cyprus	70.735	63.5	51.096	5.25	196.868
Finland	23.865	21.68106	19.766	1.619	92.231
France	28.215	24.53916	22.055	2.216	97.875
Germany	23.369	22.13789	19.324	2.05	91.375
Greece	136.992	118.3939	122.562	7.584	821.622
Ireland	110.516	120.7518	91.919	3.085	384.344
Italy	76.534	71.83957	51.223	8.1	197.784
Netherlands	32.582	29.84436	30.638	1.727	127.831
Portugal	67.183	55.75266	51.5	6.075	382.591
Spain	67.702	67.64828	43.647	4.705	207.585

Panel B. Phase 2, from 1 August 2007 to 30 April 2010, observations = 718

Country	Mean	Median	Std. Dev.	Min	Max
Austria	103.678	86.82967	50.707	39.155	239.848
Belgium	174.867	154.0847	74.996	69.645	404.419
Cyprus	752.581	903.2707	477.772	124.391	1683.682
Finland	45.963	35.70949	19.762	23.854	90.174
France	120.77	92.96382	51.134	60.136	247.309
Germany	59.533	51.01689	23.466	24.11	115.667
Greece	5037.86	3818.011	4689.62	513.692	21464.406
Ireland	517.074	573.6684	230.018	158.286	1263.406
Italy	305.116	268.8052	130.811	124.84	590.624
Netherlands	68.27	54.03103	29.195	28.405	135.452
Portugal	709.558	562.2519	354.443	198.886	1656.674
Spain	333.239	308.1213	105.668	143.947	633.486
United Kingdom	65.313	65.33173	16.213	27.835	103.562

Panel C. Phase 3, from 3 May 2010 to 29 March 2013, observations = 760

Country	Mean	Median	Std. Dev.	Min	Max
Austria	29.736	27.977	6.364	19.996	45.608
Belgium	46.282	43.816	10.448	30.769	79.415
Cyprus	559.066	433.510	286.416	243.238	1356.544
Finland	22.569	22.259	3.043	17.061	32.471
France	46.466	43.765	14.946	23.68	82.587
Germany	20.321	20.242	6.145	11.479	37.053
Greece	1127.01	977.334	748.784	376.968	5622.042
Ireland	80.16	53.588	43.285	36.255	188.826
Italy	143.574	115.273	59.746	81.038	300.402
Netherlands	29.922	28.861	13.49	13.93	58.92
Portugal	234.944	179.417	111.271	105.783	547.749
Spain	123.696	91.111	64.694	54.372	295.584
United Kingdom	24.882	20.323	9.399	15.183	51.945

Panel D. Phase 4, from 1 April 2013 to 31 December 2015, observations = 715

Country	Mean	Median	Std. Dev.	Min	Max
Austria	53.859	34.264	54.442	1.493	268.879
Belgium	74.329	45.512	78.66	1.926	404.419
Cyprus	384.327	186.977	431.465	5.25	1683.682
Finland	25.865	22.966	21.003	1.083	92.231
France	54.768	42.868	53.664	1.47	247.309
Germany	29.034	22.690	26.292	1.292	115.667
Greece	1774.48	532.026	3290.981	4.717	21464.406
Ireland	198.79	117.388	243.772	1.667	1263.406
Italy	147.535	109.996	135.552	5.291	590.624
Netherlands	36.528	31.706	32.583	1.129	135.452
Portugal	284.065	157.163	344.094	3.864	1656.674
Spain	146.71	91.981	142.229	2.348	633.486

Panel E. Full sample, from 3 October 2005 to 31 December 2015, observations = 2,670

Table A.3. Correlation matrix of the log-changes of the CDS spreads.

Country	Austria	Belgium	Cyprus	Finland	Germany	Greece	Ireland	Italy	Netherlands	Portugal	Spain
Austria	1.000										
Belgium	0.371	1.000									
Cyprus	0.014	-0.107	1.000								
Finland	0.011	0.016	-0.187	1.000							
Germany	0.105	0.103	-0.006	0.082	1.000						
Greece	0.092	0.128	-0.059	0.096	0.265	1.000					
Ireland	0.031	0.102	-0.136	0.112	-0.019	0.082	1.000				
Italy	0.069	0.183	0.050	0.050	0.153	0.335	0.041	1.000			
Netherlands	-0.201	0.027	-0.248	0.172	0.074	0.003	0.179	0.063	1.000		
Portugal	0.099	0.138	0.063	-0.018	0.168	0.275	0.071	0.304	-0.046	1.000	
Spain	0.061	0.037	0.026	0.021	0.132	0.093	0.054	0.080	0.000	0.134	1.000

Panel A. Phase 1, from 3 October 2005 to 31 July 2007, observations = 417

Country	Austria	Belgium	Cyprus	Finland	Germany	Greece	Ireland	Italy	Netherlands	Portugal	Spain
Austria	1.000										
Belgium	0.671	1.000									
Cyprus	0.229	0.254	1.000								
Finland	0.514	0.519	0.159	1.000							
Germany	0.646	0.619	0.237	0.499	1.000						
Greece	0.613	0.617	0.241	0.435	0.533	1.000					
Ireland	0.599	0.618	0.212	0.413	0.520	0.543	1.000				
Italy	0.718	0.707	0.214	0.557	0.625	0.742	0.591	1.000			
Netherlands	0.596	0.581	0.233	0.516	0.581	0.431	0.502	0.607	1.000		
Portugal	0.631	0.679	0.194	0.434	0.579	0.664	0.580	0.725	0.486	1.000	
Spain	0.665	0.732	0.256	0.468	0.609	0.662	0.606	0.758	0.574	0.734	1.000

Panel B. Phase 2, from 1 August 2007 to 30 April 2010, observations = 718

Country	Austria	Belgium	Cyprus	Finland	Germany	Greece	Ireland	Italy	Netherlands	Portugal	Spain	UK
Austria	1.000											
Belgium	0.776	1.000										
Cyprus	0.073	0.085	1.000									
Finland	0.726	0.719	0.118	1.000								
Germany	0.763	0.746	0.109	0.724	1.000							
Greece	0.249	0.261	-0.049	0.243	0.210	1.000						
Ireland	0.628	0.689	0.106	0.594	0.610	0.329	1.000					
Italy	0.707	0.793	0.123	0.668	0.695	0.302	0.780	1.000				
Netherlands	0.768	0.774	0.089	0.721	0.781	0.240	0.627	0.707	1.000			
Portugal	0.559	0.641	0.106	0.542	0.556	0.339	0.816	0.742	0.572	1.000		
Spain	0.693	0.788	0.142	0.664	0.667	0.302	0.793	0.913	0.701	0.767	1.000	
UK	0.738	0.761	0.069	0.688	0.762	0.241	0.632	0.704	0.753	0.584	0.682	1.000

Panel C. Phase 3, from 3 May 2010 to 29 March 2013, observations = 760

Country	Austria	Belgium	Cyprus	Finland	Germany	Greece	Ireland	Italy	Netherlands	Portugal	Spain	UK
Austria	1.000											
Belgium	0.501	1.000										
Cyprus	0.042	0.131	1.000									
Finland	0.335	0.353	0.035	1.000								
Germany	0.406	0.509	0.035	0.335	1.000							
Greece	0.206	0.238	0.131	0.117	0.162	1.000						
Ireland	0.405	0.580	0.186	0.303	0.415	0.362	1.000					
Italy	0.316	0.541	0.160	0.264	0.365	0.445	0.703	1.000				
Netherlands	0.328	0.343	0.059	0.250	0.328	0.177	0.329	0.238	1.000			
Portugal	0.347	0.496	0.151	0.254	0.359	0.416	0.645	0.809	0.270	1.000		
Spain	0.309	0.537	0.157	0.256	0.366	0.441	0.690	0.924	0.224	0.797	1.000	
UK	0.247	0.334	0.027	0.269	0.334	0.161	0.338	0.289	0.160	0.276	0.290	1.000

Panel D. Phase 4, from 1 April 2013 to 31 December 2015, observations = 715

Country	Austria	Belgium	Cyprus	Finland	Germany	Greece	Ireland	Italy	Netherlands	Portugal	Spain
Austria	1.000										
Belgium	0.674	1.000									
Cyprus	0.161	0.181	1.000								
Finland	0.426	0.438	0.090	1.000							
Germany	0.601	0.598	0.171	0.424	1.000						
Greece	0.322	0.338	0.095	0.214	0.283	1.000					
Ireland	0.485	0.540	0.137	0.343	0.421	0.321	1.000				
Italy	0.594	0.668	0.155	0.406	0.543	0.440	0.550	1.000			
Netherlands	0.504	0.543	0.140	0.440	0.525	0.232	0.441	0.493	1.000		
Portugal	0.528	0.603	0.155	0.329	0.495	0.431	0.554	0.718	0.406	1.000	
Spain	0.567	0.665	0.187	0.368	0.528	0.414	0.557	0.818	0.471	0.722	1.000

Panel E. Full sample, from 3 October 2005 to 31 December 2015, observations = 2,670

Table A.4. Country codes for European countries

Austria	AT	Belgium	BE	Cyprus	CY	Finland	FI
France	FR	Germany	DE	Greece	EL	Ireland	IR
Italy	IT	Netherlands	NL	Portugal	PT	Spain	ES
United Kingdom	UK						

The list of country code is from Eurostat.

REFERENCES

Alter, Adrian; Beyer, Andreas (2014): The dynamics of spillover effects during the European sovereign debt turmoil. In *Journal of Banking & Finance* 42, pp. 134–153.

Angelini, Paolo; Grande, Giuseppe; Panetta, Fabio (2014): The negative feedback loop between banks and sovereigns.

Arghyrou, Michael G.; Kontonikas, Alexandros (2012): The EMU sovereign-debt crisis: Fundamentals, expectations and contagion. In *Journal of International Financial Markets, Institutions and Money* 22 (4), pp. 658–677.

Augustin, Patrick (2014): Sovereign credit default swap premia. In *Journal of Investment Management*.

Baldwin, Richard; Beck, Thorsten; et al (2015): Rebooting the eurozone: Step 1-agreeing a crisis narrative. In *CEPR Policy Insight* No.85.

Baldwin, Richard; Giavazzi Francesco (2015): The eurozone crisis: A consensus view of the causes and a few possible remedies: CEPR.

Bampinas, Georgios; Panagiotidis, Theodore; Politsidis, Panagiotis (2020): Sovereign bond and CDS market contagion: A story from the Eurozone crisis.

Bartlett, William; Prica, Ivana (2017): Interdependence between core and peripheries of the European economy: Secular stagnation and growth in the western balkans. In *The European journal of comparative economics*.

Beirne, John; Caporale, Guglielmo Maria; Schulze-Ghattas, Marianne; Spagnolo, Nicola (2013): Volatility spillovers and contagion from mature to emerging stock markets. In *Review of International Economics* 21 (5), pp. 1060–1075.

Beirne, John; Fratzscher, Marcel (2013): The pricing of sovereign risk and contagion during the European sovereign debt crisis. In *Journal of International Money and Finance* 34, pp. 60–82.

Bhanot, Karan; Burns, Natasha; Hunter, Delroy; Williams, Michael (2012): Was there contagion in eurozone sovereign bond markets during the Greek debt crisis.

Broto, Carmen; Pérez-Quirós, Gabriel (2015): Disentangling contagion among sovereign CDS spreads during the European debt crisis. In *Journal of Empirical Finance* 32, pp. 165–179.

Bruyckere, Valerie de; Gerhardt, Maria; Schepens, Glenn; Vander Vennet, Rudi (2013): Bank/sovereign risk spillovers in the European debt crisis. In *Journal of Banking & Finance* 37 (12), pp. 4793–4809.

Buchholz, Manuel; Tonzer, Lena (2016): Sovereign credit risk co-movements in the eurozone: Simple interdependence or contagion? In *International Finance* 19 (3), pp. 246–268.

Caporin, Massimiliano; Pelizzon, Loriana; Ravazzolo, Francesco; Rigobon, Roberto (2018): Measuring sovereign contagion in Europe. In *Journal of Financial Stability* 34, pp. 150–181.

Chen, Ruo; Milesi-Ferretti, Gian Maria; Tressel, Thierry (2013): External imbalances in the eurozone. In *Economic Policy* 28 (73), pp. 101–142.

Cheng, Lu; Guo, Ruocheng; Moraffah, Raha; Sheth, Paras; Candan, K. Selcuk; Liu, Huan (2022): Evaluation methods and measures for causal learning algorithms. In *IEEE Transactions on Artificial Intelligence*.

Claeys, Peter; Vašíček, Bořek (2014): Measuring bilateral spillover and testing contagion on sovereign bond markets in Europe. In *Journal of Banking & Finance* 46, pp. 151–165.

Constâncio, Vítor (2013): The European crisis and the role of the financial system, updated on 3/10/2021, checked on 3/10/2021.

Cripps, Francis; Izurieta, Alex; Singh, Ajit (2011): Global imbalances, under-consumption and over-borrowing: The State of the world economy and future policies. In *Development and Change* 42 (1), pp. 228–261.

Croci, Elisabetta Angelini; Farina, Francesco; Valentini, Enzo (2016): Contagion across eurozone's sovereign spreads and the core-periphery divide. In *Empirica* 43 (1), pp. 197–213.

Cronin, David; Flavin, Thomas J.; Sheenan, Lisa (2016): Contagion in eurozone sovereign bond markets? The good, the bad and the ugly. In *Economics Letters* 143, pp. 5–8.

Di Quirico, Roberto (2010): Italy and the global economic crisis. In *Bulletin of Italian Politics* 2 (2).

Dornbusch, R.; Park, Y. C.; Claessens, S. (2000): Contagion: Understanding how it spreads. In *The World Bank Research Observer* 15 (2), pp. 177–197.

Eichengreen, Barry; Gupta, Poonam (2018): Managing sudden stops. In *Central Banking, Analysis, and Economic Policies Book Series* 25, pp. 9–47.

Frankel, Jeffrey (2015): The euro crisis: Where to from here? In *Journal of Policy Modeling* 37 (3), pp. 428–444.

Geffner, Hector; Dechter, Rina; Halpern, Joseph Y. (Eds.) (2022): Probabilistic and causal inference. The works of Judea Pearl. Association for Computing Machinery (ACM books, #36).

Glover, Brent; Richards-Shubik, Seth: Contagion in the European sovereign debt crisis. In *NBER Working Paper*.

Glymour, Clark; Zhang, Kun; Spirtes, Peter (2019): Review of causal discovery methods based on graphical models. In *Frontiers Genetics* 10, p. 524.

Gómez-Puig, Marta; Sosvilla-Rivero, Simón (2014): Causality and contagion in EMU sovereign debt markets. In *International Review of Economics & Finance* 33, pp. 12–27.

Gómez-Puig, Marta; Sosvilla-Rivero, Simón (2016): Causes and hazards of the euro area sovereign debt crisis: Pure and fundamentals-based contagion. In *Economic Modelling* 56, pp. 133–147.

Granger, C. W. J. (1969): Investigating causal relations by econometric models and cross-spectral methods. In *Econometrica* 37 (3), p. 424.

Grauwe, Paul de; Ji, Yuemei (2013): Self-fulfilling crises in the eurozone: An empirical test. In *Journal of International Money and Finance* 34, pp. 15–36.

Halbert White; Karim Chalak; Xun Lu (2011): Linking granger causality and the Pearl causal model with settable systems. In *NIPS Mini-Symposium on Causality in Time Series*, pp. 1–29.

He, Zonglu; Maekawa, Koichi (2001): On spurious Granger causality. In *Economics Letters* 73 (3), pp. 307–313.

Higgins, Matthew; Klitgaard, Thomas (2014): The balance of payments crisis in the euro area periphery. In *Current Issues in Economics and Finance* 20.

Hobza, Alexander; Zeugner, Stefan (2014): Current accounts and financial flows in the euro area. In *Journal of International Money and Finance* 48 (Part B), pp. 291–313.

Horváth, Bálint L.; Huizinga, Harry; Ioannidou, Vasso (2015): Determinants and valuation effects of the home bias in European banks' sovereign debt portfolios.

Hsiao, Cheng (1981): Autoregressive modelling and money-income causality detection. In *Journal of Monetary Economics* 7 (1), pp. 85–106.

Kalainathan, Diviyan; Goudet, Olivier (2019): Causal discovery toolbox: Uncover causal relationships in Python.

Kalbaska, A.; Gątkowski, M. (2012): Eurozone sovereign contagion: Evidence from the CDS market (2005–2010). In *Journal of Economic Behavior & Organization* 83 (3), pp. 657–673.

Kaminsky, Graciela L.; Reinhart, Carmen M.; Végh, Carlos A. (2003): The unholy trinity of financial contagion. In *Journal of Economic Perspectives* 17 (4), pp. 51–74.

Koutmos, Dimitrios (2018): Interdependencies between CDS spreads in the European Union: Is Greece the black sheep or black swan? In *Annals of Operations Research* 266 (1-2), pp. 441–498.

Lane, Philip (2011): The Irish crisis. In The Euro Area and the Financial Crisis.

Lane, Philip (2012): The European sovereign debt crisis. In *Journal of Economic Perspectives* 26 (3), pp. 49–68.

Longstaff, Francis A. (2010): The subprime credit crisis and contagion in financial markets. In *Journal of Financial Economics* 97 (3), pp. 436–450.

Masson, Paul R. (1998): Contagion: Monsoonal effects, spillovers, and jumps between

multiple equilibria. In *IMF Working Papers* 98 (142), p. 1.

McKinnon, Ronald I.; Pill, Huw (1998): International overborrowing: A decomposition of credit and currency risks. In *World Development* 26 (7), pp. 1267–1282.

Merler, Silvia; Pisani-Ferry, Jean (2012): Who's afraid of sovereign bonds? Brussels: Bruegel (Bruegel Policy Contribution, 2012/02).

Mink, Mark; Haan, Jakob de (2013): Contagion during the Greek sovereign debt crisis. In *Journal of International Money and Finance* 34, pp. 102–113.

Missio, Sebastian; Watzka, Sebastian (2011): Financial contagion and the European debt crisis, 9/1/2011.

Nagayasu, Jun (2001): Currency crisis and contagion: evidence from exchange rates and sectoral stock indices of the Philippines and Thailand. In *Journal of Asian Economics* 12 (4), pp. 529–546.

Nauta, Meike; Bucur, Doina; Seifert, Christin (2019): Causal discovery with attention-based convolutional neural networks. In *Machine learning and knowledge extraction* 1 (1), pp. 312–340.

Orsi, Roberto (2013): The quiet collapse of the Italian economy. Available online at https://blogs.lse.ac.uk/eurocrisispress/2013/04/23/the-quiet-collapse-of-the-italian-economy/, updated on 7/4/2015, checked on 7/8/2021.

Peet, John; La Guardia, Anton (2014): Unhappy union. How the euro crisis - and Europe - can be fixed. New York: PublicAffairs.

Pereira, Paulo T.; Wemans, Lara (2015): Portugal and the global financial crisis: short-sighted politics, deteriorating public finances and the bailout imperative.

Pfaff, Bernhard; Stigler, Matthieu (2021): VAR Modelling. Package ‘vars’.

Quaglia, Lucia; Royo, Sebastián (2015): Banks and the political economy of the sovereign debt crisis in Italy and Spain. In *Review of International Political Economy* 22 (3), pp. 485–507.

Reis, Ricardo (2015): Looking for a success in the euro crisis adjustment programs: The case of Portugal. In *Brookings Papers on Economic Activity* 2015 (2), pp. 433–458.

Rigobon, Roberto (2019): Contagion, spillover, and interdependence. In *Economía* 19 (2), pp. 69–100.

Romano, Simone (2021): The 2011 crisis in Italy: A story of deep-rooted (and still unresolved) economic and political weaknesses.

Saka, Orkun (2020): Domestic banks as lightning rods? Home bias and information during the eurozone crisis. In *Journal of Money, Credit and Banking* 52 (S1), pp. 273–305.

Santis, Roberto A. de (2012): The euro area sovereign debt crisis: Safe haven, credit rating agencies and the spread of the fever from Greece, Ireland and Portugal.

Schölkopf, Bernhard (2022): Causality for machine learning. In Probabilistic and causal inference. The works of Judea Pearl, vol. 27. First Edition, pp. 765–804.

Uribe, Martín (2006): On overborrowing. In *American Economic Review* 96 (2), pp. 417–421.

Véron, Nicolas (2007): Is Europe ready for a major banking crisis? In *Policy Briefs* (234).

Wanna, John; Lindquist, Evert; Vries, Jouke de (Eds.) (2015): Ireland's economic crisis: the good, the bad and the ugly: Edward Elgar Publishing.

Whelan, Karl (2014): Ireland's economic crisis: The good, the bad and the ugly. In *Journal of Macroeconomics* 39, pp. 424–440.

Zenios, Stavros A. (2013): The Cyprus debt: Perfect crisis and a way forward. In *Cyprus Economic Policy Review.*

CHAPTER 3

DEVELOPING EARLY WARNING SYSTEMS FOR FINANCIAL CRISIS USING MACHINE LEARNING METHODS

3.1 Introduction

Since the creation of the first modern stock trading market in Amsterdam in 1611, there has been a long history of financial crisis (Aliber and Kindleberger, 2015; Reinhart and Rogoff, 2011). Financial crises not only damage the domestic economy of the crisis country, but can also transmit shocks through contagion and spillovers to other economies. Nowadays, it is not surprising to see global financial crisis in the interconnected economies. A recent example is the 2008-2009 global financial crisis, originated from the U.S. subprime crisis, it spread over to the European continent and triggered the eurozone crisis from 2010 to 2013.

With the level of integration in today's global economies, it is imperative to establish Early Warning Systems (hereafter EWS) for financial crisis. An EWS can detect impending financial crisis and give warning information to the authorities. It helps policy makers to prevent or prepare for a potential financial crisis in sufficient time, thus to avoid loss and harm to the economy and individuals. The first EWS is proposed by Kaminsky et el. (1998), who use an indicator signaling approach to predict financial crisis. Since then, there has been a growing literature on designing new EWS.

Essentially, an EWS is a prediction mechanism which uses current and past information about the economy to predict whether there will be a crisis in the near future. In chapter 1, we have given an overview of the powerful and flexible ML algorithms. An EWS can be viewed

as a supervised learning task. Given the historical data on predictors (indicators) and labeled outcomes (dummy variable on crisis), an EWS can learn from the past and generate into unseen data. The ML methods have demonstrated their premium power in such prediction problems, making them the perfect candidate for building an EWS. There were some studies that applied ML algorithms for EWS in the past few years, but this area remained under-explored.

This chapter aims to construct EWS using machine learning (hereafter ML) methods for global financial crisis. Specifically, we want to show the strength of XGBoost (Extreme Gradient Boosting). In existing EWS literatures, the popular ML options are random forest, support vector machines and neural network, etc., XGBoost is seldomly used to establish an EWS. In applied ML, XGBoost is a popular algorithm that has been used on many real-world classification and regression problems (Nobre and Neves, 2019). XGBoost has been the champion of major competitions including the Kaggle Competitions (Chen and Guestrin, 2016). It has received rave from the ML practitioners, especially in analyzing and forecasting the stock market (Gumus and Kiran, 2017). However, in academia, there are very few relevant studies that use XGBoost. In some studies, XGBoost has been proven to be less useful than other ML methods. For example, Bluwstein et al. (2020) show that XGBoost perform less well to the other tree-based methods (e.g., random forest) in predicting financial crisis. In this chapter, our goal is to show the excellent predicting capability of XGBoost, benchmarking with random forest and the conventional logit model.

By using XGBoost and random forest on the same set of indicators (predictors), we can also rank the feature importance of the indicators. Random forest has the build in function to obtain Gini coefficient based on impurity measures. We also calculate the Shapley values, a measure borrowed from the game theory literature, to identify the predictors that are guiding the prediction. Although feature importance is not causal importance, it can still provide valuable information on crisis cycles and linkages.

Besides using the novel XGBoost algorithm, we have also assembled a crisis dataset, extended from the Jordà-Schularick-Taylor Macrohistory database (Jordà et al. 2017), by referencing several other well-known crisis datasets. Our dataset contains annual data of 17 advanced economies from 1870 to 2016. We use three models to predict crisis using the dataset: two ML methods, random forest and XGBoost, and the traditional logit model.

The remaining of this chapter proceeds as follows: section 3.2 reviews the related literature, section 3.3 presents the dataset, section 3.4 describes the three models used in this paper, section 3.5 shows prediction results, section 3.6 presents the variable significance, section 3.7 concludes.

3.2 Related literatures

There are three streams of studies that are relevant to the topic of this chapter. The first group is studies and reports on crisis datasets. The second group is development and findings in the EWS. The third one is recent ML applications in the EWS.

First, we will look at the recent crisis datasets. The crisis datasets differ in many aspects in terms of region, crisis types, frequency, etc. The most comprehensive dataset is the Global Crises Data by Country maintained by Carmen Reinhart and her colleagues at Harvard University. The Global Crises Data includes over 70 countries from 1800 to present on an annual basis, it has binary variables representing different types of crises, such as sovereign crisis, currency crisis, inflation crisis and banking crisis. This dataset is widely used in the global crisis literature. Reinhart and Rogoff (2011) used an early version of this dataset to provide a quantitative overview of financial crisis over 8 centuries.

Laeven and Valencia (2020) present the IMF database on systemic banking crises, which records 151 systemic banking crises episodes around the globe from 1970 to 2017. The authors

focus on the banking crisis in modern times. The dataset's previous versions (Laeven and Valencia 2013) and current version are used in many crises research, such as Schularick and Taylor (2012) and Laeven and Valencia (2018). Other global crisis datasets include Frankel and Saravelos (2012), who construct a dataset consist of 50 annual macroeconomic and financial variables for 96 crises.

Compared to global datasets, there are more regional crisis datasets that are available. Most of the regional crisis datasets are about advanced economies. For example, Babecký et al. (2012) develop a quarterly dataset for economic crises in the European Union (EU) and OECD countries from 1970 to 2010. The European Central Bank (ECB) has recently developed the Macro-prudential Database that record the financial crises in European countries (Lo Duca, 2017). The OECD constructs a monthly dataset that shows the recession indicators for the eurozone countries from 1960 to 2021.

Another dataset is the Jordà-Schularick-Taylor Macrohistory database (Jordà et al., 2017). "It is the result of an extensive data collection effort over several years. In one place it brings together macroeconomic data that previously had been dispersed across a variety of sources." It is the main dataset used in this paper, which contains 44 macroeconomic and financial indicators and a binary financial crisis variable for 18 advanced economics for 147 years. The researchers have used their dataset for several empirical papers, such as Jordà et al. (2016) and Jordà et al. (2018). The dataset is also used by many other economists, such as Bluwstein et al. (2020) and Tölö (2020).

The above datasets are the ones that are frequently used in the existing literature. Next, we will examine the second stream of papers on the development and findings in the EWS. The early EWS papers mainly use two approaches, the signaling approach and dependent regression analysis (logit or probit regression). Almost all the earliest literature on EWS takes the signaling approach. (Kaminsky et el. (1998) is the pioneer). In this approach, economists construct

indicators and the thresholds for such indicators. If an indicator surpasses its threshold, this equals to a warning signal for potential crisis. Papers adopting the signaling approach include Berg and Pattillo (1999), Cooper et al. (2000) and others. The early works set the stage for the regression-based models. With the indicators selected by prior works, the regression models use a binary variable as the outcome variable of a regression (indicative of a crisis), then regress it on potential macroeconomic and financial indicators. Many papers follow this path, to name a few, Borio and Lowe (2002), Reinhart and Rogoff (2007), Borio and Drehmann (2009), Schularick and Taylor (2012), Laeven and Valencia (2013), and Babecký et al. (2013). Demirgüç-Kunt and Detragiache (2005) survey the early works on signaling approach and regression analysis, the authors conclude that the logit regression models are more suitable for an EWS. Davis and Karim (2008) assess the logit regression and the signal extraction approach, their findings suggest that the logit model is the most appropriate for global EWS, while the signaling approach works better with country specific EWS.

Recent development in the regression approach is to use multinomial logit model instead of binomial models. For example, Bussiere and Fratzscher (2006) propose a multinomial logit model with a post-crisis bias component and find evidence that it outperforms binomial models. The notion that the multinomial logit model can outperform binomial models in predicting systemic banking crises is supported in later works (e.g., Caggiano et al., 2016).

These previous studies provide an abundance of early warning indicators that are proven to have practical prediction power in real-world applications. Tölö et al. (2018) construct a summary table of early warning indicators used in previous papers. Given the large pool of indicators, researchers start to introduce the predictive power of ML methods into the EWS literature.

An early example is Ahn et al. (2011), they use support vector machine (SVM) to study the Korean financial market and find that the SVM outperforms the logit model. More recent works

are listed below. Alessi and Detken (2018) use random forest to predict banking crisis with an emphasis in excessive credit growth using the quarterly dataset of EU and OECD countries constructed by Babecký et al. (2012), their early warning tree highlights similar indicators as in the earlier signaling EWS literature. Beutel et al. (2019) employ K-nearest neighbor (KNN), decision trees, random forests and other ML methods to predict the banking crisis in 15 advanced economies from 1970 to 2016 (their data is from Leaven (2018)). Even though the ML algorithms can obtain good in-sample fit, they perform poorly out-of-sample, compared to the conventional logit model.

Bluwstein et al. (2020) use extreme trees, random forest, neural networks and other machine learning methods to predict financial crisis using the Jordà-Schularick-Taylor Macrohistory database, they find that almost all the ML models outperform the logit model, especially the extreme trees and random forest, but XGBoost performs less well compared with other ML methods. Coulombe et al. (2020) take a somewhat different approach, they first design experiments to identify the "treatment" effect of different ML features, through the experiments, they study the usefulness of the underlying features that drive ML gains over conventional methods. They employ kernel ridge regression and random forest to study the FRED-MD database, which is a monthly macroeconomic database with an indicator for recession period in the U.S. Their results show that ML's ability to detect non-linearity in the data is the game-changer to the EWS literature.

Jarmulska (2020) uses random forest and logit model to study the fiscal stress events in 43 advanced economies from 1992 to 2018. The comparison between the two models suggests a clear advantage of the ML methods in forecasting. Tölö (2020) employs LSTM and GRU neural nets that use the representation power of deep learning algorithms on the Jordà-Schularick-Taylor Macrohistory database, their novel ML methods outperform basic neural networks and the logit model.

In this line of research to employ ML methods in EWS, only very few studies have used XGBoost and recognized its importance. Chatzis et al. (2018) build EWS to forecast stock market crisis using daily stock, bond and currency data from 39 countries, they employed ML methods such as support vector machines, random forest, XGBoost and deep neural networks. The authors claim that they are the first one to apply XGBoost and deep learning in the context of financial crisis forecasting. Their results show that all the ML methods demonstrate excellent predictive power, especially the deep neural networks. Huang (2020) applies logit regression, random forest and XGBoost on Germany's credit default records, the ML methods perform better than the logit model, XGBoost has reached about 80% accuracy.

In most of the studies that adopt ML methods in macroeconomic forecasting, researchers find clear evidence to support the use of ML. The only exception is Beutel et al. (2019), in which the authors find the logit model out performs ML methods in out-of-sample fit. Among all the ML methods that are employed, the most commonly used and praised ML algorithm is random forest, whereas XGBoost is rarely used. In the following sections, we will show that XGBoost can perform just as well as random forest in building an EWS.

3.3 Data description

The Jordà-Schularick-Taylor Macrohistory database (hereafter JST dataset) is a comprehensive and harmonized dataset on 17 advanced economies from 1870 to 2016, it includes a binary variable for systemic financial crisis, and 44 macroeconomic and financial variables such as GDP, import and export, and equity dividend return. The 17 countries are Australia, Belgium, Canada, Denmark, Finland, France, Germany, Italy, Japan, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom and United States (see Jordà et al., (2017) for details on the dataset).

Comparing to other crisis datasets, the systemic financial crisis binary variable in the JST dataset shows some differences, some crisis periods in the JST datasets are categorized as non-crisis periods in other datasets, some non-crisis periods are marked as crisis period in other datasets. Also, out of the 2499 observations, there are only 90 crisis periods (3.6%), such a low incidence rate is much smaller than most other datasets. Therefore, this paper extends the systemic financial crisis binary variable in the JST dataset by referencing other crisis datasets.

The crisis variable used in this paper is mostly based on the Global Crises Data by Country maintained (hereafter GCD) by Carmen Reinhart and her colleagues. The crises in the GCD are categorized into 4 groups, which are banking crisis, balance of payment crisis, sovereign crisis and inflation crisis. Because our paper aims to establish an EWS for financial crisis using ML methods, and the three types of crises (banking crisis, balance of payment crisis, sovereign crisis) are collectively recognized as financial crisis (Laeven and Valencia, 2008). Therefore, we dropped the inflation crisis binary variable in the GCD and merged the three binary variables for banking crisis, balance of payment crisis and sovereign crisis into one binary variable for financial crisis.

In almost all cases, when there is an inflation crisis in the GCD, the crisis country simultaneously has a sovereign crisis or balance of payment crisis. In rare cases when there is a standalone inflation crisis, we turn to other crisis datasets to determine whether it is a financial crisis. For all crisis incidents, we cross-examine all available datasets to refine the financial crisis binary variable so that it is consistent with the majority of the datasets.

In the EWS literature, there are two ways to treat the crisis binary variable for prediction problems. One way is to predict the whole crisis periods with forward looking, i.e., the crisis dummy is indicative of the actual crisis periods, the indicators are lagged 2 years with regard to the crisis dummy (e.g., Jarmulska (2020)). The other way is to predict the crisis one or two years prior to a crisis, the crisis dummy is in fact a signal dummy. The actual crisis periods are

excluded from the data (e.g., Bluwstein et al. (2020)). We take the first approach because of our data structure. The JST dataset is highly imbalanced (only 3.6% of all observations are crisis periods), even after we extended the crisis dummy in the JST dataset, there are still too many non-crisis periods versus the crisis periods (only 20% of the data are marked as crisis period). If the ongoing crisis periods are all excluded as described in the second approach, the dataset will be even more imbalanced (only 9% of the data have the positive value of one). Therefore, we follow the first approach by keeping all the crisis periods.

Another treatment of the dataset is to exclude the extreme periods, this is a common practice in the EWS literature (Bluwstein et al., 2020). the JST dataset spans from 1870 to 2016, which encompasses the two world wars and the Great Depression during the 1930s. During such extreme economic and political times, financial crises happened in most countries, therefore, the period of 1914-1918 (The World War I), 1933-1939 (The Great Depression) and 1939-1945 (The World War II) are excluded from the data. Also, following the first approach, a two-year lag is introduced to the dataset, so that the crisis can be predicted in advance. For example, for crises in 2010-2012, we use the indicators in 2008-2010 as explanatory variables. The final adjustment to the dataset is to clear out the observations with missing values in one or more macroeconomic variables. Given all the adjustments, the final dataset contains 1570 observations for 17 countries, 322 out of the 1570 observations (20%) are crisis periods.

Next, we will discuss the macroeconomic and financial variables in the dataset. There are many indicators that have been proven to be predicative in previous studies (e.g., Tölö et al., 2018). However, because the JST dataset goes back to as far as 1870, it is very difficult to find data in the 19th century. Therefore, the explanatory variable selection is restricted to the variables in the original JST dataset.

Following previous studies that also use the JST dataset (e.g., Bluwstein et al., 2020; Tölö, 2020), we pick out 15 explanatory variables, which can be put into three categories: domestic

economy, competitiveness and global economy. The 13 variables in the domestic and competitiveness categories are all listed in the table of "survey of Early Warning Indicators" in Tölö (2018). They are widely used in previous studies and at least have shown some predictive power. For the two variables in the global economy category, they are derived from credit and yield curve in the domestic economy category, the global credit is the mean of all other countries' credits, the global yield curve is the mean of all other countries' yield curves.

The first category, domestic economy, includes 10 variables. They represent the macroeconomic fundamentals of an economy. The nominal GDP and GDP per capita measure the overall output of an economy. A large GDP decline is an indicator for debt crises (Babecký et al., 2014). The reason that we include both GDP and GDP per capita is because GDP is highly correlated with other variables, therefore GDP is excluded from the logit model. CPI is the adjusted domestic consumer price index in U.S. dollars, the growth rate of CPI measures the domestic inflation level, high inflation is an indicator for sovereign crisis and balance of payment crisis (Christofides et al., 2016). Money is the domestic broad money supply, it is one of the monetary policies that can be used to adjust the interest rate, an increase in money supply is a sign of injecting credit into the market in times of economic downturn (Reinhart et al., 1998). Consumption and investment are key indicators of aggregated economic activities, they directly reflect the contraction or expansion of an economy. Credit is the total loans to non-financial private sector created through the private banks, and controlled by the central banks. Credit boom periods tend to be followed by unusually low returns to equities (Davis and Taylor, 2019).

The yield curve is defined as the domestic long-term interest rate subtracting the short-term interest rate, this indicator is the same as Bluwstein et al. (2020), the authors show that the yield curve is a key indicator for financial crisis prediction, that the yield curve often steepen at the onset of a recession. Public debt is the growth rate of debt to GDP ratio. High levels of public

debt usually signal a deteriorating fiscal condition and a potential sovereign crisis. The debt service ratio is the product of credit and long-term interest rate, then divided by nominal GDP. It measures the economy's ability to repay its private debts, a high debt service ratio suggests vulnerability in the banking sector. This simple measure of debt service ratio follows Bluwstein et al. (2020), it is constructed only with the data in the JST dataset due to data availability, thus it cannot reflect features such as short-term lending rates or the maturity structure of the debt.

The second category is competitiveness, which stands for a country's competitiveness in both goods and services in international trade. The current account is represented by the growth rate of the current account to GDP ratio, which measures a country's earnings and spending abroad. Sustained large current account deficit is a sign of loss of competitiveness that might lead to a balance of payment crisis. The level of current account deficits is also robustly associated with the severity of crises (Babecký et al., 2013). Export is an important measure of a country's overall performance on the global markets, export growth suggests a comparative advantage and an export decline implies a loss of competitiveness. Import is not included here, because it is collinear with export and current account deficits. The USD exchange rate is an important indicator for balance of payment crisis, exchange rate can interact with domestic and foreign prices to determine the capability of a country's export and import.

The third category is the global economy. We follow Bluwstein et al. (2020) and Jarmulska (2020) to include two global indicators: the global yield curve, and the global credit. They are calculated by averaging the values of all other countries. Those two global indicators can represent the cross-border contagion and spillovers in the global market. The limitation of the two indicators is that they only represent the 17 advanced economies in the JST datasets, so the shocks from other countries (e.g., shocks from Greece during the eurozone crisis) cannot be picked up.

Table 3.1. Explanatory variables summary

	Description
Domestic economy	
GDP	GDP (nominal, local currency), y-o-y growth
GDP per capita	Real GDP per capita (PPP), y-o-y growth
CPI	Consumer prices (index, 1990=100), y-o-y growth
Money	Broad money (nominal, local currency), y-o-y growth
Consumption	Real consumption per capita (index, 2006=100), y-o-y growth
Investment	Investment-to-GDP ratio, y-o-y growth
Credit	Total loans to non-financial private sector (nominal, local currency), y-o-y growth
Yield curve	Long-term interest rate - Short-term interest rate, in levels
Public debt	Public debt-to-GDP ratio, y-o-y growth
Debt service ratio	Credit × long-term interest rate over GDP, y-o-y growth
Competitiveness	
Current account	Current account-to-GDP ratio, y-o-y growth
Export	Exports (nominal, local currency), y-o-y growth
USD exchange rate	USD exchange rate (local currency/USD), y-o-y growth
Global economy	
Global yield curve	Mean of all other countries' yield curves
Global credit	Mean of all other countries' credit

Explanatory variables used in this paper, all data is from the JST dataset, y-o-y growth is the year-to-year growth rate in basis points.

Table 3.1 explains the makeup and structure of the variables. 13 variables use the growth rate, 2 variables (yield curve and global yield curve) are in levels. A two-year lag is introduced in the dataset. Table 3.2 shows the descriptive statistics of the variables for the crisis subgroup and the non-crisis subgroup. A t-test is performed to check the difference in mean in those variables, 9 out of the 15 variables have a significant difference in their means.

Table 3.2. Descriptive statistics of the explanatory variables

	Crisis Mean	Crisis Std. Dev.	Non-crisis Mean	Non-crisis Std. Dev.	Difference in mean
Domestic economy					
GDP	6.507	7.929	7.109	7.658	-0.602
GDP per capita	1.52	3.672	2.525	3.215	-1.005***
CPI	4.203	6.238	3.236	5.175	0.966***
Money	0.771	3.754	0.409	3.059	0.361*
Consumption	1.241	3.766	2.363	3.331	-1.122***
Investment	-0.066	2.293	0.1	1.66	-0.166
Credit	9.651	10.02	9.557	8.621	0.094
Yield curve	0.509	2.219	0.734	1.74	-0.224*
Public debt	1.151	5.362	-0.244	5.019	1.395***
Debt service ratio	4.455	3.161	3.488	2.523	0.967***
Competitiveness					
Current account	-0.022	2.636	0.064	2.158	-0.086
Export	9.013	21.337	10.149	40.366	-1.136
USD exchange rate	2.405	16.552	0.527	9.619	1.878***
Global economy					
Global yield curve	0.74	0.969	.813	.874	-0.073
Global credit	9.259	6.17	10.097	5.379	-0.837**

The difference in mean t-test results. Significant levels are $*p<0.1$; $**p<0.05$; $***p<0.01$.

Figure 3.1 shows the result of a Principal Coordinate Analysis (PCoA) of the data for crisis and non-crisis subgroups, the variability in the whole data set is not negligible as the points spread out over the horizontal coordinate, but the points for crisis and non-crisis subgroups display little dissimilarity between them.

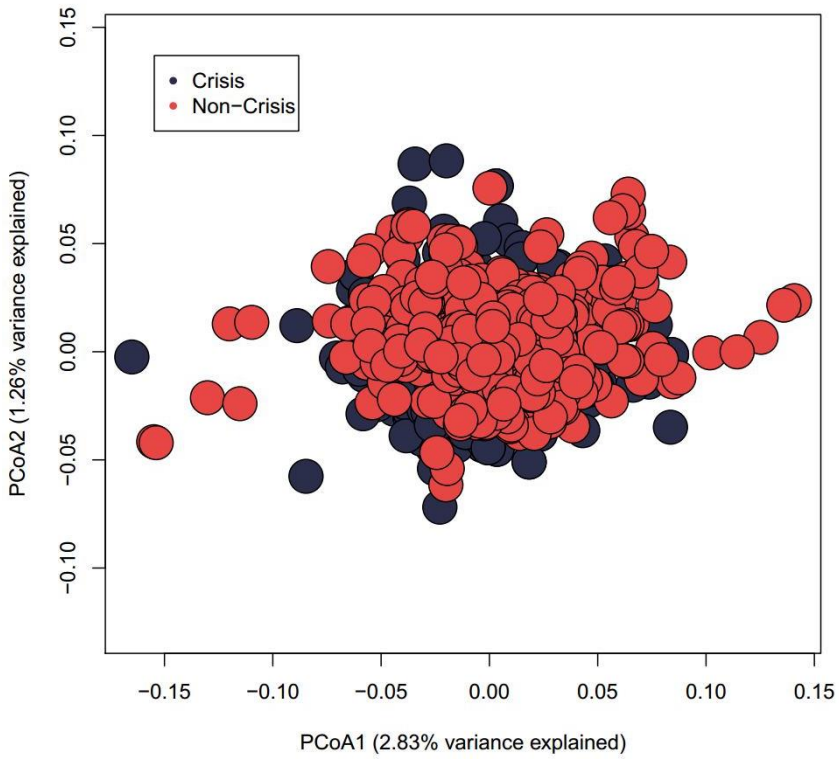

Figure 3.1. Principal Coordinate Analysis (PCoA) of the data for crisis and non-crisis subgroups.

The correlation matrix of the explanatory variables is shown in Figure 3.2. GDP is the only variable that exhibits high correlation with other variables. This is because some variables are calculated by the growth rate of its ratio to the GDP. Therefore, we perform a VIF (variable inflation factor) to measure the amount of multicollinearity among the variables. As expected, the GDP has a VIF above 9, suggesting high collinearity. The two ML methods (random forest and XGBoost) are naturally immune to multicollinearity, because the algorithms only pick one of the collinear variables when deciding a split at a node in the trees, the multicollinearity does not affect prediction performance. But with the detection of multicollinearity, we will report results of the algorithms both with the GDP and without GDP. In the benchmark logit model, the GDP is excluded from the regression.

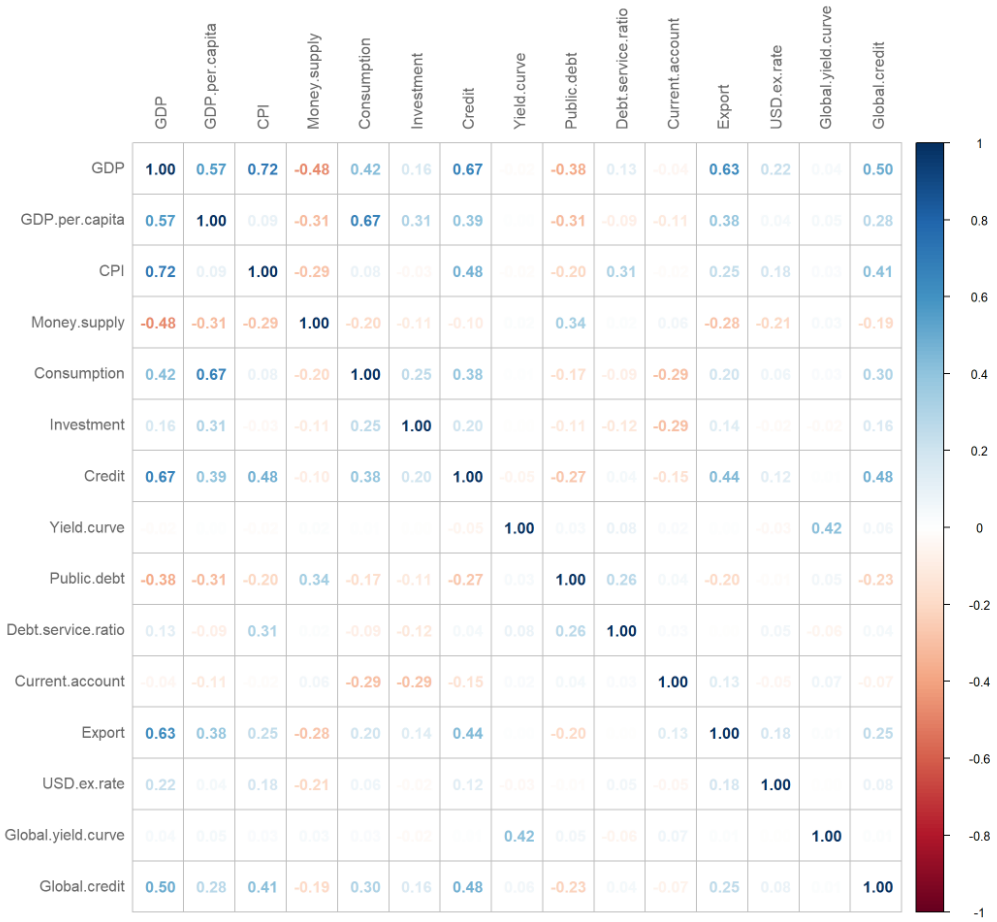

Figure 3.2. Correlations matrix of the explanatory variables. Blue means a positive correlation, red means negative. The magnitude of the correlation is shown by color intensity, the darker the color, the larger the absolute value.

3.4 Methodology

This chapter uses three models to establish the EWS, XGBoost, random forest, and the benchmark logit model. The first one is XGBoost, as discussed in section 3.1, XGBoost is a very popular method in applied ML because of its impressive results in real-world applications, specifically in finance. XGBoost combines the advantages of gradient boosting with random forest, making it a powerful and efficient tool for prediction problems. The EWS literature has only seen very few XGBoost applications. The second model is random forest. Previous studies

have demonstrated its outstanding prediction performance in crisis detection, random forest also outperforms a lot of other ML methods such as decision trees and support vector machines. Besides the two ML methods, the conventional logit model is used as the benchmark model. Logit regression is the standard model of the EWS in crisis prediction. It has been shown to outperform many ML algorithms including random forest (Beutel et al., 2019).

This paper aims to illustrate the predictive power of XGBoost in establishing EWS, we will compare the performance of the three models in the following sections. The next section presents the three models in detail.

3.4.1 XGBoost

XGBoost is a recent development in ML built upon decision trees and gradient boosting (Chen and Guestrin, 2016). XGBoost is a tree-based ensemble ML algorithm which uses a gradient boosting framework. It is best suited to smaller panel data (in ML, panel data is also referred as tabular data) for a variety of applications such as regression and classification.

XGBoost improves over the previous algorithms through parallel processing, tree pruning and handling missing values and regularizations. XGBoost algorithms starts out with building the base tree predictions using parallelized implementation. After the trees are built, XGBoost specifies the maximum depth of each tree and prune the trees backwards. XGBoost also has a built-in regularization component that penalizes more complex models through both LASSO and ridge to prevent overfitting.

In a XGBoost framework, given a dataset with N observations, the explanatory variables x_i has m features so that $x_i \in R^m$. For each x_i, there is an outcome variable y_i, $y_i \in R$. In the classification problems, $y_i \in (0, 1)$. First, a tree ensemble model is performed to predict the outcome variable $\hat{y}_i$ using the explanatory variables x_i and the following K additive functions

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \tag{3.1}$$

where f_k corresponds to an independent tree structure with leaf weights, F is the tree space. The goal of the regularized learning is to minimize equation 3.2

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i,\ y_i) + \sum_i \Omega(f_k) \tag{3.2}$$

where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and y_i. Ω is an additional regularization term for penalizing the complexity of the model.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega_i\|^2 \tag{3.3}$$

where T is the number of leaves, ω_i is the weight of the i^{th} leaf. Define I_j as the instance set of leaf j. The optimal values for ω_j can be obtained by solving equation 3.1 to 3.5.

$$\omega_j^* = -\frac{\sum_{i\in I_j} \partial_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i^{t-1}\right)}{\sum_{i\in I_j} \partial^2{}_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i^{t-1}\right) + \lambda} \tag{3.4}$$

$$\tilde{\mathcal{L}}^t(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} \partial_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i{}^{t-1}\right)\right)^2}{\sum_{i\in I_j} \partial^2{}_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i{}^{t-1}\right) + \lambda} + \gamma T \tag{3.5}$$

Normally it is impossible to enumerate all the possible tree structures. Assume that $I = I_L \cup I_R$ where I_L and I_R are the instance sets of child left and right nodes. The following formula is usually used in practice.

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L}\partial_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i{}^{t-1}\right)\right)^2}{\sum_{i\in I_L}\partial^2{}_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i^{t-1}\right)+\lambda} + \frac{\left(\sum_{i\in I_R}\partial_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i{}^{t-1}\right)\right)^2}{\sum_{i\in I_L}\partial^2{}_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i^{t-1}\right)+\lambda} - \frac{\left(\sum_{i\in I}\partial_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i{}^{t-1}\right)\right)^2}{\sum_{i\in I}\partial^2{}_{\hat{y}^{t-1}} l\left(y_i,\ \hat{y}_i^{t-1}\right)+\lambda}\right] - \gamma \tag{3.6}$$

More details about the XGBoost algorithm can be found in the authors' paper (Chen and Guestrin, 2016).

3.4.2 Random forest

Random forest is a combination of decision tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. None of the trees in the forest can see the entire data, therefore avoiding the problem of overfitting (Breiman, 2001). In a random forest algorithm, the data is recursively split into partitions, at each non-terminal node in the forest, the split is done by asking a question with binary answers, the answer then determines the questions that will be asked at the next node. This process repeats until the data reaches the terminal node, where a categorical outcome is produced. The criterion for splitting the data at each node is based on impurity measures such as Gini impurity and entropy.

Here, we will use the Gini impurity, which is a function of measuring the quality of split in each node, to stratify the predictor space during the recursive binary splitting process. The Gini impurity for node N is defined as

$$g(N) = \sum_{i\neq j} P(\omega_i)P(\omega_j) \tag{3.7}$$

where ω_i is the weight for the i^{th} leaf, $P(\omega_i)$ is the proportion of the population of class i in node N.

The goal of random forest is to minimize the impurity by choosing the best split in each node. The best split is thus defined as the highest reduction in impurity or the highest gain in information. The information gain of choosing a split is defined as

$$\Delta I(N) = I(N) - P_L \times I(N_L) - P_R \times I(N_R) \tag{3.8}$$

where the decrease in the impurity measure $I(N)$ equals the current level of $I(N)$ minus the expectation of the two child nodes of node N. P_L is the proportion of data in node N that goes to the left child node, correspondingly, P_N is the proportion of data in node N that goes to the right child node, $I(N_L)$ and $I(N_R)$ are the impurity measure for the two child nodes.

The above optimization problem at each node set the threshold values for the trees in the random forest. Random forest uses the ensemble methods of bagging (as known as Bootstrap aggregating) to aggregate the decision trees predictors, which is often associated with large variance and low out-of-sample prediction accuracy. To sum up, random forest is a nonparametric way to estimate the set of outcome variables y_i from a set of explanatory variables x_i.

3.4.3 Logit model

In the binomial logit model, the outcome variable Π_i for observation i is predicted from a vector of covariates $x_{1i}, \cdots, x_{mi}$. The logit model is described as below

$$\Pi_i = \frac{e^{\beta + \beta_1 x_{1i} + \cdots + \beta_m x_{mi}}}{1 + e^{\beta + \beta_1 x_{1i} + \cdots + \beta_m x_{mi}}} = \frac{1}{1 + e^{-(\beta + \beta_1 x_{1i} + \cdots + \beta_m x_{mi})}} \tag{3.9}$$

where Π_i is the probability of a crisis period, x_{ki} is the value of k^{th}covariates, k is from 1 to m. β_i is the coefficient of each individual covariate. The value of β_i is estimated using the method of maximal likelihood.

3.4.4 Performance measure

AUROC curve (areas under the receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters, the True Positive Rate (TPR) on the vertical axis and the False Positive Rate (FPR) on the horizontal axis. The area under the ROC curve (AUROC) is a common measure used in the EWS literature. All recent works have adopted this measure in their model assessments (e.g., Tölö, 2020; Jarmulska, 2020). Therefore, we also use the AUROC to evaluate the three models under study.

For any thresholds, the two parameters in a ROC curve are defined as

$$TPR = \frac{Ture\ positive}{Ture\ positive + False\ negative} \tag{3.10}$$

$$FPR = \frac{Ture\ negative}{False\ positive + True\ negative} \tag{3.11}$$

The area under the ROC curve can be interpreted as the probability that the distribution of thresholds during the crisis is stochastically larger than during normal times (Drehmann and Juselius, 2014). Therefore, the AUROC provides a convenient and interpretable measure of the EWSs.

The values of AUROC lie between 0 and 1. A pure random prediction which assigns the sample into 2 groups would result in an average AUROC of 0.5. When the AUROC is below

0.5, it suggests that the model is uninformative, when the value is above 0.5, the model is informative, when the value is 1, it is fully informative.

3.5 Results

This part shows the AUROC of the three models. For XGBoost and random forest, two AUROCs are reported, one is when GDP is included in the explanatory variables, one is without. For the logit model, only one AUROC is reported without GDP as an explanatory variable.

Figure 3.3 and Figure 3.4 show the AUROC for XGBoost when GDP is included, and not included, respectively. Both the AUROC have a value above 0.5, the first one has a value of 0.812, while the second one has a value of 0.994, which is almost a fully informative model.

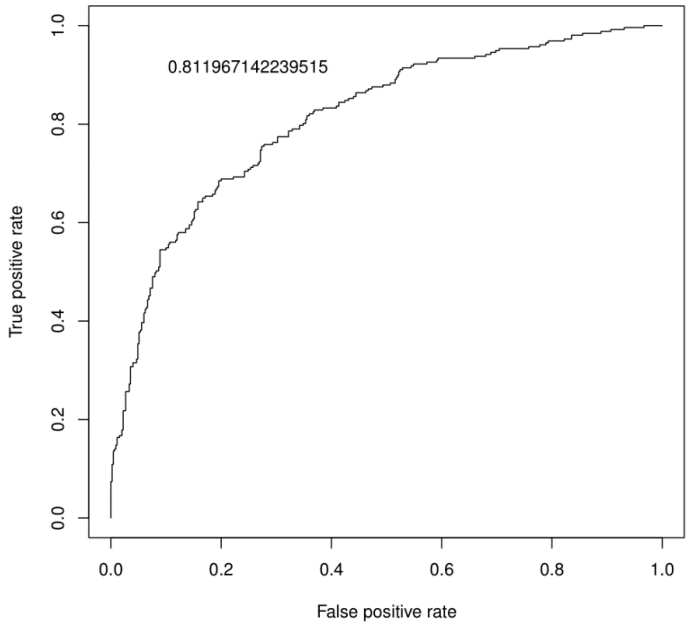

Figure 3.3. AUROC for XGBoost with GDP included in the explanatory variables.

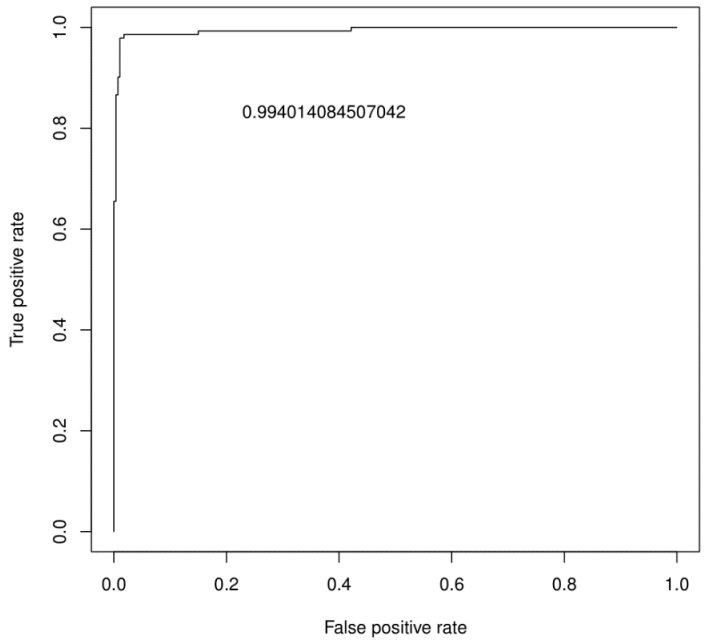

Figure 3.4. AUROC for XGBoost with GDP excluded in the explanatory variables.

Figure 3.5 and Figure 3.6 show the AUROC for random forest when GDP is included, and not included, respectively. Both the AUROC have a value above 0.5, the first one has a value of 0.809, while the second one has a value of 1, which is a fully informative model.

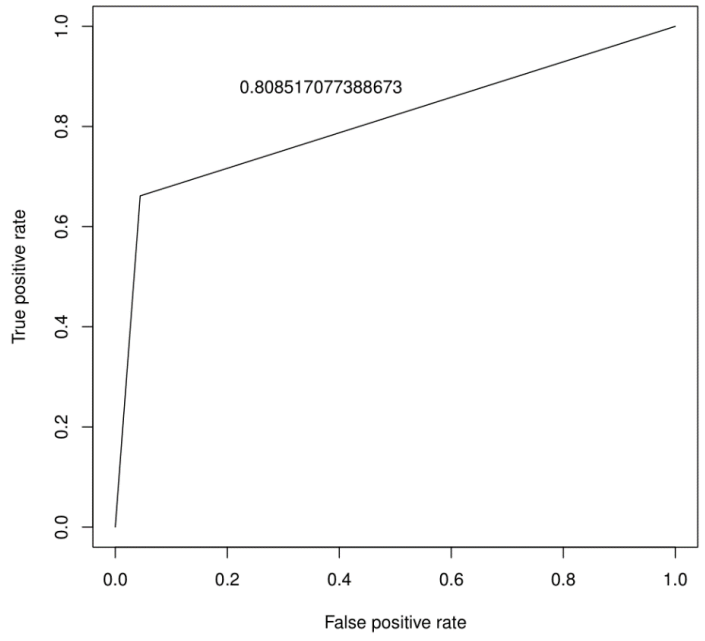

Figure 3.5. AUROC for random forest with GDP included in the explanatory variables.

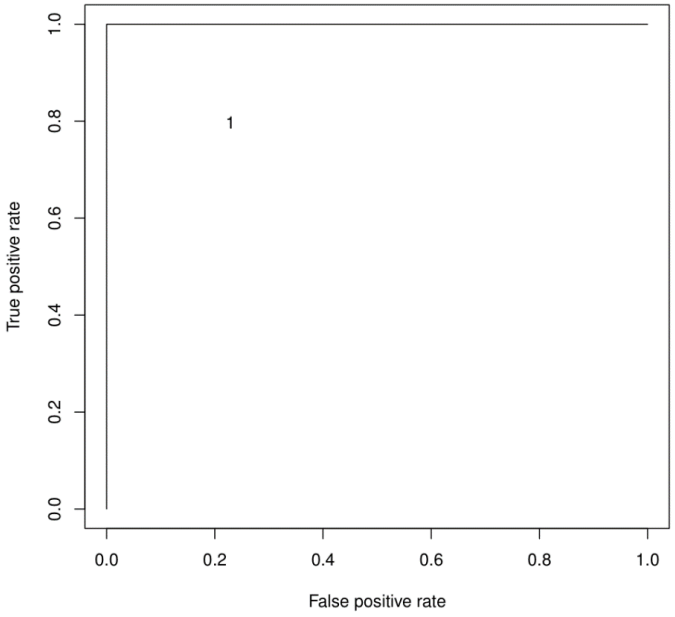

Figure 3.6. AUROC for random forest with GDP excluded in the explanatory variables.

Figure 3.7 shows the AUROC for the logit model with GDP excluded in the explanatory variables. It has a value of 0.724 which is above 0.5; the model is informative. Table 3.3 summarizes the finding of the AUROCs.

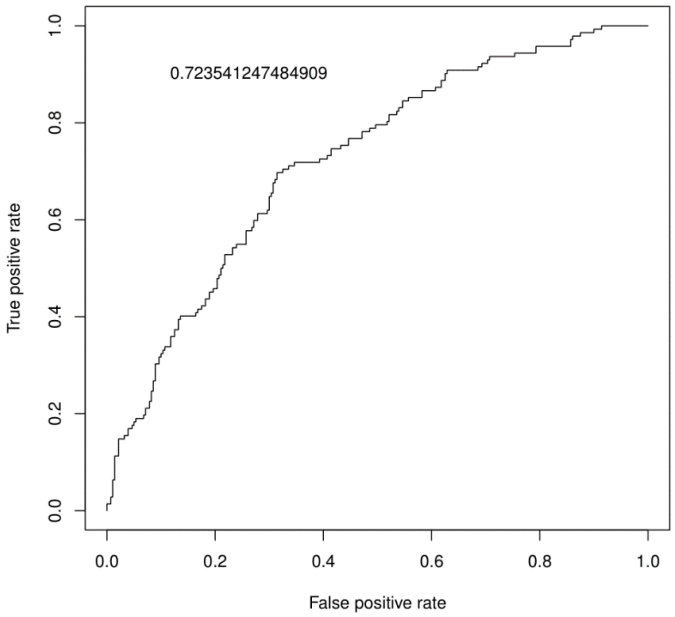

Figure 3.7. AUROC for the logit model with GDP excluded in the explanatory variables.

Table 3.3. AUROCs for the three models.

Model	AUROC with GDP	AUROC without GDP
Random Forests	0.809	1
XGBoost	0.812	0.994
Logit		0.724

The results of the AUROC show that the ML algorithms perform extremely well in the test set. The XGBoost and random forest have reached a value greater than 0.99, while the benchmark logit model has a value of 0.724. This finding shows that the ML methods outperform the logit model in terms of AUROC.

In theory, the multicollinearity associated with the GDP should have little effect on the predictive performance of XGBoost and random forest. However, as shown in Table 3.3, the AUROC is higher in the predictions without GDP in both XGBoost and random forest, suggesting that GDP should be removed from the explanatory variables given this specific dataset. Between XGBoost and random forest, the values of AUROC are almost identical, showing equal predictive power in this regard.

3.6 Variable significance

In an EWS, besides prediction accuracy, another important task is to understand the importance of indicators and their thresholds. The indicators are not only important in forecasting crisis, they can also point out potential areas of vulnerability of an economy. This section uses the Shapley values borrowed from the game theory literature, which can identify the variables that are guiding the prediction.

The Shapley value, introduced in Shapley (1953), is a method for assigning payouts to players depending on their contribution to the total payout in a coalition game. In ML prediction tasks, the Shapley value of a variable is the average of all the marginal contributions of this variable to all possible coalitions with other explanatory variables, that is the difference between the actual prediction of a given observation and the average of other "coalition" predictions with other explanatory variables (Jarmulska, 2020). The Shapley value is an ideal measure for variable importance in the ML prediction tasks (Molnar, 2019), the existing literature has used the Shapley value extensively as a criterion for feature importance (e.g., Bluwstein et al., 2020; Jarmulska, 2020; Tölö et al., 2020). The Shapley value can interpret a model's prediction with regard to attribution of the various features, dependence on the feature's value, and the most important features (Lopez de Prado, 2020). For detailed information about how the Shapley values are formulated, please see Strumbelj and Kononenko (2014).

Rather than using a typical bar chart to rank the values, this paper reports the Shapley values density scatter plot for each explanatory variable. The plot can show how much impact each variable has on the prediction in the test set. Due to package incompatibility, the Shapley plot for XGBoost differs from that of random forest and the logit model, the difference will be explained below.

The Shapley summary plot of the XGBoost reports the Shapley values by a scatter plot, which is the distribution of contributions for each explanatory variable. The mean of each variable's Shapley value is reported in numbers and ranked from the largest to the smallest. The color of each dot represents the level of impact of the variable on each observation in the test set (Lundberg and Lee, 2017). Figure 3.8 shows that the two most influential variables in XGBoost are the global yield curve and CPI. The global yield curve affects almost all the predictions in the test set with similar impact. The yield curve, ranking 3rd on the list, effects a few predictions with a larger impact.

Figure 3.8. Shapley summary plot for XGBoost, GDP is excluded from the model.

Figure B.1 in Appendix B shows the Shapley summary plot for XGBoost when GDP is included. In theory, including the GDP should have minor effects on the XGBoost's prediction. But in practice, the Shapley values change a lot just as the AUROCs. The 2 most influential variables now become the global yield curve and the domestic yield curve, both of them have a sizeable impact, followed by consumption and CPI.

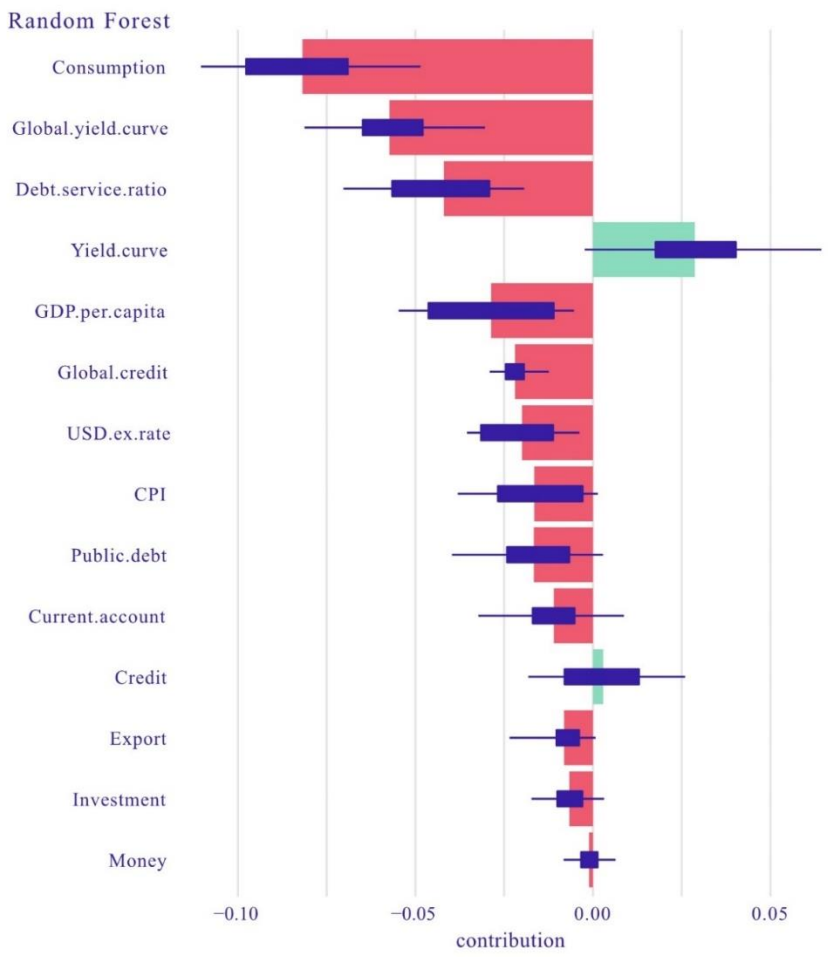

Figure 3.9. Shapley summary plot for random forest, GDP is excluded from the model.

The second graph (Figure 3.9) is the Shapley plot for random forests when GDP is excluded from the explanatory variables. In this plot, the red and green bars show the means of the Shapley values for each variable, the box plots summarize the distribution of contributions for each explanatory variable (Biecek and Burzykowski, 2020). The list of explanatory variables is sorted by their means (from the largest to the smallest). In Figure 3.9, the two most impactful variables are consumption and the global yield curve. Almost all the variables have negative effects on the prediction, while the domestic yield curve and credit have a positive effect. The domestic yield curve ranks 4th in the list.

The Shapley summary plot of the random forest with GDP included in the model is reported in the Appendix. Figure B.2 shows that consumption and GDP per capita are the two most impactful variables. The global yield curve now ranks 6th on the list, though it affects some of the predictions, its effect is minor since its average Shapley value is around 0. The influence of the global yield curve becomes positive in this figure. Two other variables, CPI and the current account, also turn positive. It is interesting to see such dissimilar results in variable importance by introducing a collinear variable.

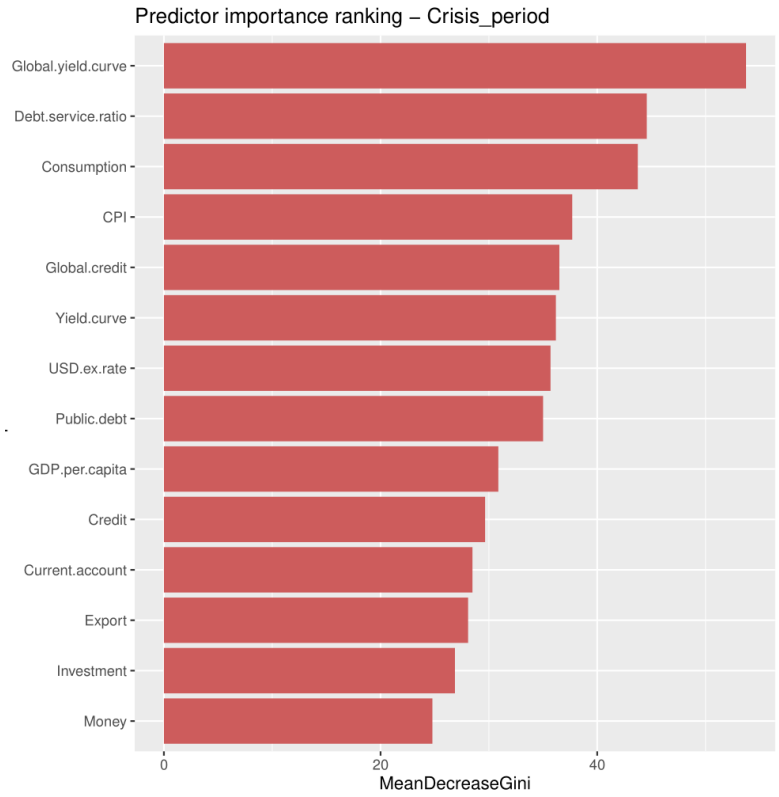

Figure 3.10. Gini index for random forest, excluding GDP.

Another variable importance metric is the Gini-based importance (Gini index) obtained by training the random forest algorithm. For each variable, its Gini impurity decreases every time that variable is chosen to split a node across every tree of the forest. The accumulated decrease

is divided by the number of trees in the forest to give an average, which is the value of the Gini index. The larger the Gini index, the more important the variable is (the scale of the Gini index is irrelevant). In Figure 3.10, the two most important variables are global yield curve and debt service ratio.

The third Shapley summary plot is for the logit model excluding the GDP. This plot has the same attributes as Figure 3.9. In Figure 3.11, the two most influential variables are consumption and GDP per capita. In order to compare the Shapley value results with the traditional variable significance, the logit regression coefficients are also reported in Table 3.4.

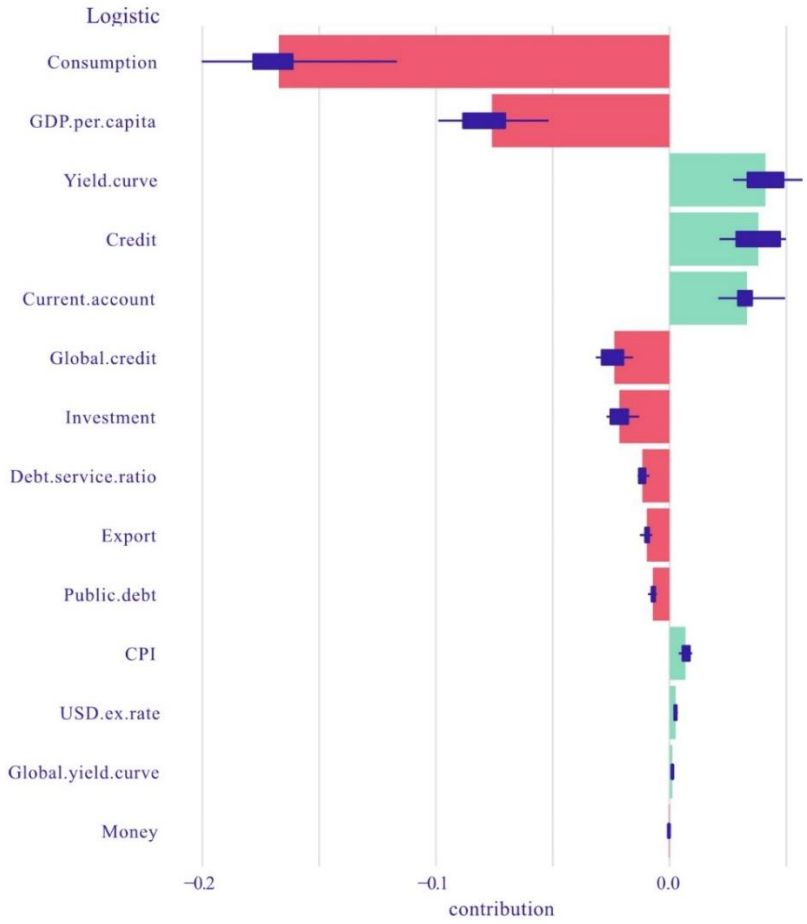

Figure 3.11. Shapley summary plot for logit model, GDP is excluded from the model.

Table 3.4. Logit regression results

Variable	Estimate	Std.Error	z statistic	Pr(>|z|)
GDP per capita	-9.829	6.624	-1.484	0.137
CPI	7.180	5.737	1.251	0.210
Money	2.089	4.841	0.432	0.666
Consumption	-24.923	6.618	-3.766	0.000***
Investment	-4.287	5.005	-0.856	0.391
Credit	11.287	6.135	1.840	0.065*
Yield curve	-10.173	4.854	-2.096	0.036**
Public debt	5.310	4.960	1.071	0.284
Debt service ratio	13.973	4.459	3.133	0.001***
Current account	-12.852	4.660	-2.758	0.005***
Export	28.467	13.026	2.185	0.028**
USD exchange rate	-3.402	4.269	-0.797	0.425
Global yield curve	0.307	4.402	0.070	0.944
Global credit	-13.107	5.258	-2.493	0.012

*z-test significant levels are * p< 0.1; ** p<0.05; *** p<0.01.*

The variables that are significant are consumption, debt service ratio, current account, export, yield curve and credit. The top six variables on the Shapley list are consumption, GDP per capita, yield curve, credit, current account, and global credit. Four of them are the same, including the most significant variable, consumption. In both measures, consumption has a large negative impact on the predictions, which is consistent with the findings in random forest (Figure 3.9).

Table 3.5 summarizes the variable importance for all the metrics discussed above. The Shapley value and the Gini index are reported with rank order while the logit coefficients are reported in value with significance level.

The two ML methods recognize the global yield curve as an important indicator for predicting a financial crisis. The domestic yield curve is also an important indicator but with a

smaller impact. The domestic yield curve is also a significant variable in the logit model. Therefore, the yield curves are the most important indicators in this dataset. Intuitively and empirically, an inverted domestic yield curve is one of the most reliable leading indicators of an impending recession (Reinhart and Rogoff, 2011). Recent studies who use the ML methods to identify the early warnings also find the yield curves as an important measure (e.g., Bluwstein et al., 2020; Tölö, 2020).

Table 3.5. Variable importance summary across the three methods

Variable	Shapley R. forests	Shapley XGBoost	Shapley Logit	Coefficient Logit	Gini index R. forests
GDP per capita	5	11	2	-9.829	9
CPI	8	2	11	7.180	4
Money	14	14	14	2.089	14
Consumption	1	5	1	-24.923***	3
Investment	13	10	7	-4.287	13
Credit	11	4	4	11.287*	10
Yield curve	4	3	3	-10.173**	6
Public debt	9	8	10	5.310	8
Debt service ratio	3	7	8	13.973***	2
Current account	10	6	5	-12.852***	11
Export	12	13	9	28.467**	12
USD ex. rate	7	9	12	-3.402	7
Global yield curve	2	1	13	0.307	1
Global credit	6	12	6	-13.107	5

Logit regression results, significant levels are $*p< 0.1$; $**p<0.05$; $***p<0.01$.

The global yield curve is a global indicator for all countries, reflecting changes in the global economies, which serves as a warning of potential contagion and spillovers during a global financial crisis. Our dataset contains only 17 advanced economies in this dataset, 13 of them are

European countries, 11 are eurozone countries. Given the economic and political interdependencies among the European countries, the domestic economic condition of a country would have effects on its neighbors. However, without further investigation, this finding should not be generalized in macroeconomic forecasting to a larger scale, due to limitation of the dataset.

Besides the yield curves, the second group of important indicators are consumption and CPI. Consumption is the most important variable in random forest and the logit model. It ranks 5th for the XGBoost. For the logit model, the importance of consumption is confirmed by both the Shapley value and the coefficient significance. Just like the yield curve, abrupt change in consumption is a classic indicator for imminent financial crisis, CPI is a measure of inflation and is often seen as the pressure indicator for overall economic health.

The three variables in the competitiveness category, current account, USD exchange rate and export have low prediction power in all three models. This could be due to the structure of the dataset, in the 19th century and the early 20th century, the volume of global international trade only makes up a small share of the total economy, and some of the countries have a peg with the U.S. dollar, hence there is not much variation in the data. Table 3.2 confirms that the difference between crisis periods and non-crisis periods in the current account and export are not significant.

A consistency exists in all three models, that money has the lowest prediction power, there could be many reasons why this is the case, one explanation is the 2-year lag in the explanatory variables. Since the money supply can be controlled by the authorities, this policy-oriented metric is not a barometer for the economic health, but rather reflects the expectation of the authorities. Another reason is the formation of the European Union. Eleven out of the 17 countries in our dataset joined the eurozone in 1999 with a unified monetary policy, therefore the money supply indicator in our dataset may not capture the variations in these economies.

3.7 Conclusion

This paper uses two popular ML algorithms to study the extended JST dataset with comparison to a benchmark logit model. Our goal is to show the predictive power of the ML algorithms, especially the XGBoost. XGBoost not only outperforms the logit model, its predictions are almost fully informative, suggesting high levels of accuracy. The performance of XGBoost is on par with the random forest. The variable importance is measured by the Shapley value in this study and the results show coherent findings with the existing literature.

Three major contributions are delivered in this study. First, we extended the JST dataset, referencing other famous crisis datasets. Second, through comparing with random forest and the logit model, we have shown that the XGBoost has excellent prediction accuracy for EWS. Between the two ML methods, random forest has been proven to have excellent predictive power in previous studies (e.g., Chatzis et al., 2018; Coulombe et al., 2020). XGBoost, which is rarely used in previous literatures, also shows outstanding prediction performance using our dataset. Third, by using the Shapley value, the variable importance can be identified. The yield curves and consumption have the most influence on the prediction, where some of these findings might only reflect the features of the JST dataset.

For future work, with the growing availability of big data in the macroeconomics, a larger dataset can help to train the algorithms and provide richer information about the economies. Also, ML is a fast-growing field, newly developed algorithms are emerging every day, so future works can better leverage the more advanced ML algorithms. Lastly, even though the early warning system for financial crisis is a pure prediction problem, causal inference could still be introduced to the framework.

Appendix B

Figure B.1. Shapley summary plot for XGBoost, GDP is included.

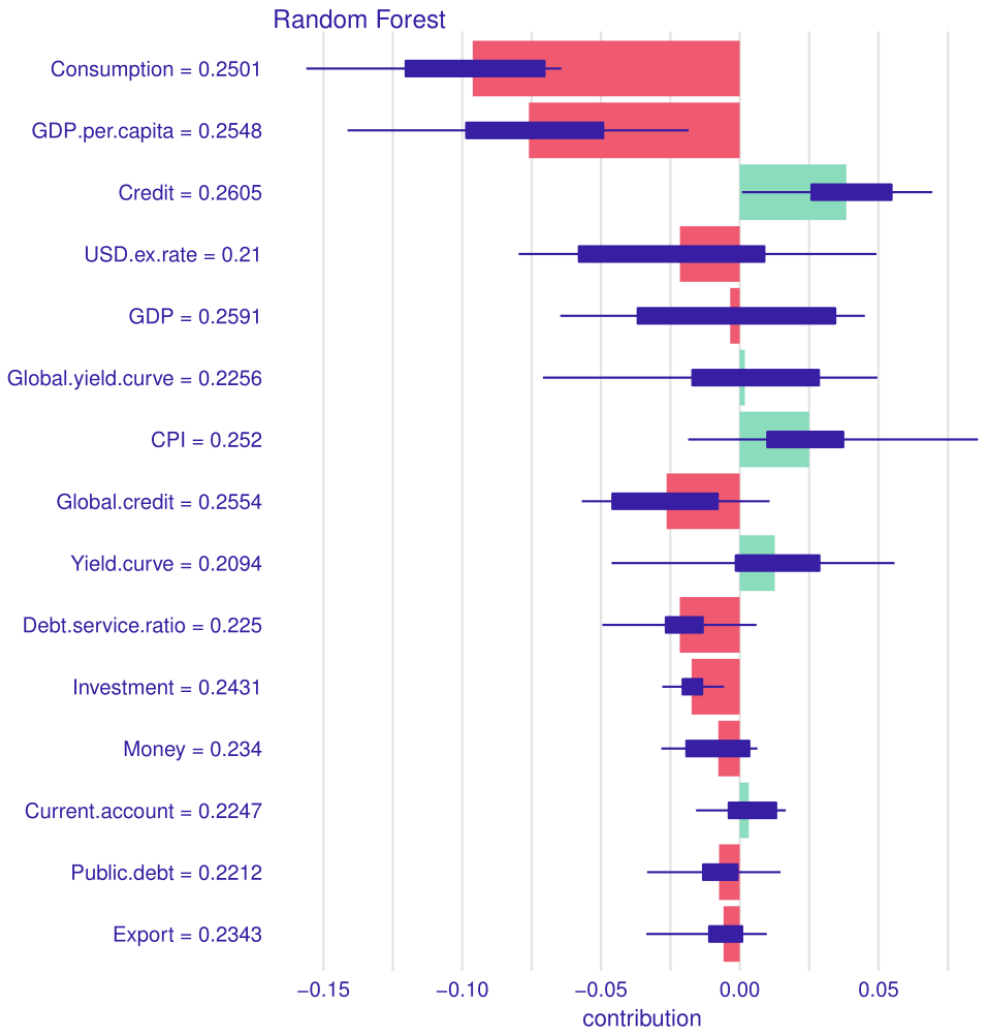

Figure B.2. Shapley summary plot for random forest, GDP included.

REFERENCES

Ahn, Jae Joon; Oh, Kyong Joo; Kim, Tae Yoon; Kim, Dong Ha (2011): Usefulness of support vector machine to develop an early warning system for financial crisis. In *Expert Systems with Applications* 38 (4), pp. 2966–2973.

Alessi, Lucia; Antunes, Antonio; Babeckk, Jan; Baltussen, Simon; Behn, Markus; Bonfim, Diana et al. (2015): Comparing different early warning systems: Results from a horse race competition among members of the macro-prudential research Network. In *SSRN Journal*.

Alessi, Lucia; Detken, Carsten (2018): Identifying excessive credit growth and leverage. In *Journal of Financial Stability* 35, pp. 215–225.

Aliber, Robert Z.; Kindleberger, Charles P. (2015): Manias, panics, and crashes. A history of financial crises, seventh edition. Seventh edition 2015. London: Palgrave Macmillan.

Babecký, Jan; Havranek, Tomas; Mateju, Jakub; Rusnák, Marek; Smidkova, Katerina; Vasicek, Borek (2012): Banking, debt and currency crises: early warning indicators for developed countries. In *ECB Working Paper Series*.

Babecký, Jan; Havránek, Tomáš; Matějů, Jakub; Rusnák, Marek; Šmídková, Kateřina; Vašíček, Bořek (2013): Leading indicators of crisis incidence: Evidence from developed countries. In *Journal of International Money and Finance* 35, pp. 1–19.

Babecký, Jan; Havránek, Tomáš; Matějů, Jakub; Rusnák, Marek; Šmídková, Kateřina; Vašíček, Bořek (2014): Banking, debt, and currency crises in developed countries: Stylized facts and early warning indicators. In *Journal of Financial Stability* 15, pp. 1–17.

Berg Andrew; Pattillo, Catherine (1999): Are currency crises predictable? A test. In *IMF Economic Review* 46 (2), pp. 107–138.

Beutel, Johannes; List, Sophia; Schweinitz, Gregor von (2019): Does machine learning help us predict banking crises? In *Journal of Financial Stability* 45, p. 100693.

Biecek, Przemyslaw; Burzykowski, Tomasz (2020): Explanatory model analysis. Explore, explain, and examine predictive models. With examples in R and Python.

Bluwstein, Kristina; Buckmann, Marcus; Joseph, Andreas; Kang, Miao; Kapadia, Sujit; Simsek, Özgür (2020): Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach. In *Bank of England working papers* (848).

Borio, Claudio; Drehmann, Mathias (2009): Assessing the risk of banking crises - revisited. In *BIS Quarterly Review*.

Borio, Claudio; Lowe, Philip (2002): Assessing the risk of banking crises. In *BIS Quarterly Review*.

Breiman, Leo (2001): Random forests. In *Machine Learning* 45 (1), pp. 5–32.

Bussiere, Matthieu; Fratzscher, Marcel (2006): Towards a new early warning system of financial crises. In *Journal of International Money and Finance* 25 (6), pp. 953–973.

Caggiano, Giovanni; Calice, Pietro; Leonida, Leone; Kapetanios, George (2016): Comparing logit-based early warning systems: Does the duration of systemic banking crises matter? In *Journal of Empirical Finance* 37, pp. 104–116.

Chatzis, Sotirios P.; Siakoulis, Vassilis; Petropoulos, Anastasios; Stavroulakis, Evangelos; Vlachogiannakis, Nikos (2018a): Forecasting stock market crisis events using deep and statistical machine learning techniques. In *Expert Systems with Applications* 112, pp. 353–371.

Chen, Tianqi; Guestrin, Carlos (2016): XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, pp. 785–794.

Christofides, Charis; Eicher, Theo S.; Papageorgiou, Chris (2016): Did established early warning signals predict the 2008 crises? In *European Economic Review* 81, pp. 103–114.

Cooper, Richard N.; Goldstein, Morris; Kaminsky, Graciela L.; Reinhart, Carmen M. (2000): Assessing financial vulnerability: An early warning system for emerging markets. In *Foreign Affairs* 79 (6), p. 176.

Coulombe, Philippe Goulet; Leroux, Maxime; Stevanovic, Dalibor; Surprenant, Stéphane. (2020): How is machine learning useful for macroeconomic forecasting?

Davis, E. Philip; Karim, Dilruba (2008): Comparing early warning systems for banking crises. In *Journal of Financial Stability* 4 (2), pp. 89–120.

Davis, Josh; Taylor, Alan (2019): The leverage factor: Credit cycles and asset returns. Cambridge, MA.

Demirgüç-Kunt, Asli; Detragiache, Enrica (2005): Cross-country empirical studies of systemic bank distress: A survey. In *National Institute of economic review* 192, pp. 68–83.

Domenico; Reichlin, Lucrezia; Small, David (2008): Nowcasting: The real-time informational content of macroeconomic data. In *Journal of Monetary Economics* 55 (4), pp. 665–676.

Drehmann, Mathias; Juselius, Mikael (2014): Evaluating early warning indicators of banking crises: Satisfying policy requirements. In *International Journal of Forecasting* 30 (3), pp. 759–780.

Erik; Kononenko, Igor (2014): Explaining prediction models and individual predictions with feature contributions. In *Knowledge and Information Systems* 41 (3), pp. 647–665.

Frankel, Jeffrey; Saravelos, George (2012): Can leading indicators assess country vulnerability? Evidence from the 2008–09 global financial crisis. In *Journal of International*

Economics 87 (2), pp. 216–231.

Gumus, Mesut; Kiran, Mustafa S. (2017): Crude oil price forecasting using XGBoost.

Jarmulska, Barbara (2020): Random forest versus logit models: which offers better early warning of fiscal stress? In *ECB Working Paper Series* (2408).

Jordà, Òscar; Schularick, Moritz; Taylor, Alan; Ward, Felix (2018): Global financial cycles and risk premiums. In *National Bureau of Economic Research*.

Jordà, Òscar; Schularick, Moritz; Taylor, Alan M. (2016): Sovereigns versus banks: credit, crises, and consequences. In *Journal of the European Economic Association* 14 (1), pp. 45–79.

Jordà, Òscar; Schularick, Moritz; Taylor, Alan M. (2017): Macrofinancial history and the new business cycle facts. In *NBER Macroeconomics Annual* 31 (1), pp. 213–263.

Junyu, Huang (2020): Prediction of financial crisis based on machine learning.

Kaminsky, Graciela; Lizondo, Saul; Reinhart, Carmen M. (1998): Leading indicators of currency crises. In *Staff Papers - International Monetary Fund* 45 (1), p. 1.

Laeven, Luc; Valencia, Fabian: Systemic banking crises: A new database. In *IMF Working Papers* (2008/224).

Laeven, Luc; Valencia, Fabian (2018): Systemic banking crises databases revisited. In *IMF Working Papers* 18 (206), p. 1.

Laeven, Luc; Valencia, Fabian (2020): Systemic banking crises database: A timely update in COVID-19 times. In *CEPR Discussion Papers* (14569).

Laeven, Luc; Valencia, Fabián (2013): Systemic banking crises database. In *IMF Economic Review* 61 (2), pp. 225–270.

Lo Duca, Marco; Koban, Anne; Basten, Marisa; Bengtsson, Elias; Klaus, Benjamin; Kusmierczyk, Piotr et al. (2017): A new database for financial crises in European countries: ECB/ESRB EU crises database. In *ECB Occasional Paper Series*.

Lopez de Prado, Marcos (2020): Interpretable machine earning: Shapley.

Lundberg, Scott M.; Lee, Su-In (2017): A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30.

Molnar, Christoph (2019): Interpretable machine learning. A guide for making black box models explainable. Available online at https://christophm.github.io/interpretable-ml-book/.

Reinhart, Carmen M.; Rogoff, Kenneth S. (2011): This time is different. Eight centuries of financial folly. Princeton University Press.

Schularick, Moritz; Taylor, Alan M. (2012): Credit booms gone bust: Monetary policy, leverage cycles, and financial crises, 1870–2008. In *American Economic Review* 102 (2), pp. 1029–1061.

Shapley, L. S. (1953): Stochastic games. In *Proceedings of the National Academy of Sciences* 39 (10), pp. 1095–1100.

Strumbelj, Erik; Kononenko, Igor (2010): An efficient explanation of individual classifications using game theory. In *The Journal of Machine Learning Research* 11, pp. 1–18.

Tölö, Eero (2020): Predicting systemic financial crises with recurrent neural networks. In *Journal of Financial Stability* 49, p. 100746.

Tölö, Eero; Laakkonen, Helinä; Kalatie, Simo (2018): Evaluating indicators for use in setting the countercyclical capital buffer. In *International Journal of Central Banking* 14 (2), pp. 51–112.