

WARSAW UNIVERSITY OF TECHNOLOGY

AUTOMATION, ELECTRONICS AND ELECTRICAL ENGINEERING

FIELD OF ENGINEERING AND TECHNICAL SCIENCES

PH.D. THESIS

Michał J. Falkowski, MSc, Eng

**Root Cause Analysis in complex multi-loop control
systems**

Supervisor

Paweł D. Domański, PhD, DSc, University Professor

Warsaw 2023

Declaration of Authorship

I, Michał J. Falkowski, MSc, Eng, declare that this thesis titled, “Root Cause Analysis in complex multi-loop control systems” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: September 27, 2023

POLITECHNIKA WARSZAWSKA

Streszczenie

Wydział Elektroniki i Technik Informatycznych
Szkoła Doktorska Politechniki Warszawskiej

Rozprawa doktorska

Analiza zależności przyczynowo-skutkowych w złożonych wielo-pętlowych systemach sterowania

mgr inż. Michał Falkowski

Właściwe, wstępne przetwarzanie i warunkowanie surowych danych procesowych jest kluczowym elementem każdej analizy rzeczywistych, wieloskalowych procesów przemysłowych. Uzyskiwane dane są zwykle niekompletne i obciążone licznymi obserwacjami oraz nieznanymi szumami. Często zauważa się wpływ czynnika ludzkiego. Powoduje to wiele nieścisłości, które mogą mieć wpływ nie tylko na samą analizę, jak również interpretację uzyskanych wyników.

Powyższe ma szczególne znaczenie w znajdowaniu błędnych obserwacji, które mogą łatwo rozprzestrzeniać się pomiędzy jednostkami procesowymi, ze względu na wzajemne powiązania przepływu materiału lub informacji. Problem wykrywania i izolowania błędów dla tych procesów jest ściśle związane z analizą przyczynowo-skutkową.

Rozwiązaniem może być zastosowanie Analizy Przyczynowości. Metody te mają na celu znalezienie związku przyczynowo-skutkowego między różnymi sygnałami i ścieżkami błędu, przy użyciu wybranych zestawów danych i/lub informacji procesowych. Ten rodzaj analizy jest używany w wielu aspektach inżynierskich i nieinżynierskich. W przypadku rozważanego obszaru inżynierskiego, analityka danych może być z powodzeniem wykorzystywana do tworzenia złożonych modeli układów przemysłowych, składających się z wielu pętli sterowania, uczestniczących w kompleksowym zadaniu sterowania. Kluczowym elementem jest znalezienie odpowiednich relacji między jego składowymi.

Kilka metod zapewnia skuteczne rozwiązanie tego problemu zakładając występowanie zależności liniowych. Niestety, rzeczywiste procesy przemysłowe wykazują właściwości nieliniowe. Znaczące i nieznanne nieliniowości ograniczają możliwe podejścia, gdyż niektóre z założeń mogą nie zostać spełnione. Rezygnacja z podejścia opartego na modelu uniezależnia rezultat metody od możliwego błędu modelu, co stawia badania po bezpiecznej stronie, stąd rozważane metody ograniczają się do tych niewymagających modelu.

W pracy przedstawiono wyniki analizy przyczynowo-skutkowej propagacji błędów regulacji z wykorzystaniem podejścia Transfer Entropy. Przygotowano kompleksową metodologię, a także pomyślnie ją zweryfikowano przy pomocy symulacyjnych i rzeczywistych danych procesowych. Pierwszym wyborem do analizy jest

podstawowa wersja Transfer Entropy. Wkład naukowy stanowi odejście od klasycznej definicji omawianej metody poprzez nietypowe podejście do funkcji rozkładu prawdopodobieństwa, będącej podstawą jej działania. Towarzyszy temu szeroko pojęta analiza danych, obejmująca proces dekompozycji na różnych poziomach.

Słowa kluczowe: Transfer Entropy, propagacja błędów, wartości odstające, przyczynowość, wielkoskalowe systemy przemysłowe, dane procesowe, metody oparte na danych

WARSAW UNIVERSITY OF TECHNOLOGY

Abstract

Faculty of Electronics and Information Technology
Doctoral School of Warsaw University of Technology

Ph.D. Thesis

Root Cause Analysis in complex multi-loop control systems

by Michał J. Falkowski, MSc, Eng

Proper pre-processing and conditioning of raw process data is a key element in any kind of analysis of real large-scale industrial processes. The acquired data is usually incomplete and affected by numerous observations and unknown noises. The influence of the human factor is also often observed. This causes many inaccuracies that may affect not only the analysis itself but also the interpretation of the obtained results.

The above is of particular importance in finding faulty observations that can easily propagate between process units due to the interconnections of material or information flows. The problem of fault detection and isolation for these processes is strictly related to the root cause analysis.

The solution may be found in the use of Causality Analysis. These methods aim at finding a root-cause relationship between different signals and fault paths, using given datasets and/or process information. This kind of analysis is used within many engineering and non-engineering contexts. In the case of the considered engineering area, data analytics can be successfully used to create complex models of industrial plants, consisting of many control loops participating in the overall process control task. The key element is to find proper relations between its elements.

Several methods provide effective solutions to this problem under the assumption of linear relationships. Unfortunately, real industrial processes exhibit nonlinear properties. Significant and unknown nonlinearities limit possible approaches, as certain assumptions might not be met. The resignation from the model-based approach makes the result independent of the possible error of the model, which puts the research on the safe side, and thus the considered methods are limited to the model-free ones.

This work presents the results of Root Cause Analysis of control errors propagation using the Transfer Entropy approach. The comprehensive methodology is prepared and successfully validated with simulation and real process data. The first choice for the analysis is the classical Transfer Entropy approach. The scientific contribution is a departure from the classical definition of the discussed method by using an unusual approach to the probability distribution function, which is the basis of its performance. This is accompanied by a broadly understood data analysis,

including the decomposition process at various levels.

Keywords: Transfer Entropy, fault propagation, outliers, causality, large-scale industrial systems, process data, data-driven methods

Acknowledgements

I would like to express my heartfelt gratitude to Grupa Azoty, Zakłady Azotowe "Puławy" SA, for invaluable support in providing the essential data required to complete this thesis. The data from the ammonia synthesis installation is instrumental in the success of this research endeavor.

Additionally, thanks to Paweł D. Domański for the invaluable substantive help during the entire research process and the preparation of this doctoral thesis.

Contents

Declaration of Authorship	iii
Streszczenie	v
Abstract	vii
Acknowledgements	ix
1 Introduction	1
2 Research rationale	3
3 The research area review	15
3.1 Topology capturing methods	20
3.1.1 Cross-correlation	20
3.1.2 Granger Causality	21
3.1.3 Partial Directed Coherence	22
3.1.4 Transfer Entropy	23
3.2 Review of parametric statistical models	25
3.2.1 Gaussian normal distribution	25
3.2.2 Robust statistics – Huber’s logistic estimator	27
3.2.3 The family of α -stable distributions	28
Cauchy probabilistic density function	29
Laplace double exponential distribution	30
t Location-scale distribution	31
3.3 Review of non-parametric statistical models	32
3.3.1 Darbellay-Vajda algorithm	33
3.3.2 Fixed-Bins algorithm	33
3.3.3 Kernel Density Estimation algorithm	34
3.4 The trend identification	34
3.5 Outliers detection	35
3.5.1 Extreme studentized deviate	36
3.5.2 InterQuartile Range method	36
3.5.3 Hampel filter	36
3.6 Data decomposition methods	37
3.6.1 Ensemble Empirical Mode Decomposition	37

3.6.2	Median Ensemble Empirical Mode Decomposition	37
4	Description of the systems	39
4.1	Description of the simulation system	39
4.2	Description of the real industrial system	48
5	Implementation of the classical Transfer Entropy method	65
5.1	Causality for the dataset with Gaussian noise	71
5.2	Causality for the dataset with Gaussian noise and Cauchy disturbance	72
6	Impact of the PDF on the Transfer Entropy method	75
6.1	Parametric statistical models	75
6.1.1	Causality for the dataset with Gaussian noise	80
6.1.2	Causality for the dataset with Gaussian noise and Cauchy disturbance	81
6.2	Non-parametric statistical models	83
6.2.1	Causality for the dataset with Gaussian noise	83
6.2.2	Causality for the dataset with Gaussian noise and Cauchy disturbance	86
7	Impact of outliers on the Transfer Entropy method	91
7.1	Causality for the dataset with Gaussian noise	95
7.2	Causality for the dataset with Gaussian noise and Cauchy disturbance	97
8	Implementation of the Transfer Entropy method for noise signals	99
8.1	Causality for the dataset with Gaussian noise	102
8.2	Causality for the dataset with Gaussian noise and Cauchy disturbance	104
9	Implementation of the Transfer Entropy method for oscillatory signals	107
9.1	Causality for the dataset with Gaussian noise	110
9.2	Causality for the dataset with Gaussian noise and Cauchy disturbance	111
10	Research overview and industrial validation	115
10.1	Summary	115
10.2	Validation with the ammonia synthesis installation dataset	120
11	Conclusion	137
	List of Figures	139
	List of Tables	145
	List of Abbreviations	147
	Physical Constants	149

*To my beloved wife Wioletta, and children, Adam, and Filip, who
were the greatest motivation to complete this thesis.*

Chapter 1

Introduction

The field of “causality” has seen much activity in recent years, both in its foundational and theoretical aspects and in industrial applications. However, it remains rare to draw the distinction (recognized by Rubin, 1974) between two different problem areas within it: assessing (in individual cases, or in general) the likely effects of applied or considered interventions - the problem of “effects of causes”, EoC; and assessing, in an individual case, whether or not an observed outcome was caused by an earlier intervention or exposure - the problem of “causes of effects”, CoE.

Philosophers have debated causality for millennia, and have propounded a large variety of conceptions and approaches. Statisticians, on the other hand, had traditionally been reluctant to imbue their inferences with causal meaning. But in recent years much more attention has been given to what exactly represents the term “causality”. Particularly influential have been the contributions of Rubin, 1974, who promoted a formulation based on “potential outcomes”, and of Pearl, 2009, based on graphical representations.

Implicit in both these approaches is the idea of a cause as an intervention applied to a system, in line with the “agency” interpretation of causality (Reichenbach, 1956; Price, 1991; Hausman, 1998; Woodward, 2003; Woodward, 2016). The main task for causality is to make inferences about the effects - that is, understanding the “effects of causes” - based on data. When making use of data, it is important to distinguish between data generated through dedicated experiments and purely observational data.

These considerations concern mainly humanities (in particular, philosophy). With the growing interest in the subject, the considerations moved to other fields of life sciences, e.g. medicine, economics, or even meteorology. The engineering sciences are also not omitted, but the nature of causality is understood in a different way. The purpose of the analysis, in this case, is to find the cause in the sense of root cause error (or fault), the phenomenon of “effect” is known or often predicted using other available methods. In engineering applications, causality is always accompanied by the concept of relationship and “correlation”. It is important to note that “correlation” does not imply “causality”. Even if two variables are found to be correlated, it does not necessarily say much about the root-cause relationship. None of the analyses would make sense without data generated artificially (simulations) or obtained from industrial plants (processes). In the second case, we are dealing with

very complex objects, resulting in a huge amount of collected data or observations. However, the real causality relations might be unclear or unknown, even though the technological structure of the process is well known.

Over the years, many methods have been developed to deal with this issue. They can be generally divided into those requiring a model of the tested object and the so-called model-free approaches. In practice, in the root cause analysis, the appropriate data-based method should be selected carefully based on known process dynamics, the available data, and the type of error. In the literature, there may be many examples of often conditional solutions, like cross-correlation analysis, Granger causality and its extensions, frequency domain methods, information-theoretical methods, and Bayesian nets. One should consider whether they are sufficient. The Transfer Entropy approach has been successfully used in simulation (theoretical) applications, but it is hard to find publications based on real industrial cases. The preliminary analysis of the topic shows the enormous potential of the chosen method in this case. Combined with the multithreading associated with data analysis, the conducted research sheds light on its hitherto unconsidered possibilities. It seems that engineers are not interested in causality, only relationships, which in fact gives no answer to the doubts that arise.

The document is organized in the following order. Chapter 2 describes the rationale of the problem and argues the legitimacy of searching for solutions to the considered problem. Chapter 3 presents selected methods, to which the authors of publications most often refer, exploring the issues of root cause analysis. Chapter 4 describes the simulation system used in calculations and presents a real industrial system as a case study example. Chapter 5 presents the results of the research carried out on simulation data, using the Transfer Entropy method in its basic version. Chapter 6 introduces the consideration of the parameters' sensitivity of the selected method on causality, more precisely, the impact of Probability Density Function fitting on the quality of the results obtained. In Chapter 7, the author focuses on the nature of the analyzed data that affects the results, and more specifically, the existence of outliers in the data. Chapter 8 is a continuation of the discussion on the nature of the data, showing the impact of the components of the analyzed signals (in this case noises) on the root cause analysis. In Chapter 9, oscillations are used as another data component for analysis. In Chapter 10, the analytical procedure is proposed, justified by an industrial example. Finally, Chapter 11 summarizes the important conclusions based on the entire research process and discusses identified open research issues.

Chapter 2

Research rationale

The detection and diagnosis of plant-wide disturbances is a major problem in the process industry. Because of the high degree of interconnections among different parts in a large-scale complex system, a simple fault may propagate along information and material flow pathways and affect other, even distort parts of the system. There are several typical reasons for poor process performance, including inappropriate controller tuning, process nonlinearity, malfunctioning of actuators and sensors, and severe mismatch in the process control (Braun et al., 2002). Faults that are commonly caused by an inappropriate control loop operation may lead to low productivity of the whole plant, can increase operational costs, and, in the most dangerous cases, may result in an unplanned plant shutdown or damage. To determine the root causes of certain anomalies, it is important to capture the process connectivity and to find the fundamental connecting pathways. Process data analysis embodies a broad viewpoint that leads to these types of problems. It is substantially different from conventional techniques, although it may not seem so at first. It generally has the benefit of providing greater insight into the problems under consideration. The structured and explicit nature of the process data analysis is very helpful in bringing to light information that can be highly useful in the decision-making of a given system. Efficient techniques for process analysis and controller assessment are important for identifying areas for further process and control improvement.

Proper process data analysis strongly depends on key process variables, which are often interpreted differently among disciplines, datasets, and research contexts. Signal selection defines the task, and further, impacts potential results. Selecting key variables and interactions is, therefore, an important step in achieving more accurate predictions, better interpretation, and identification of key subgroups for further fault analysis. In large-scale complex systems, the separation of crucial variables for the analysis becomes the problem itself. There is also a factor related to the nature of the data - it is sampled from industrial systems that are in most cases nonlinear, non-stationary, of complex origin, and contaminated by noise. To overcome these adversities, there are a lot of fault detection research methods that may be used.

The fault detection problem has been an active field of research for the past decades, because of the ever-increasing demand for higher performance, safety, and reliability standards. The considerations mainly focused on the optimization of fault detection systems (Ding et al., 2000; Gao and Ding, 2007; Yin, Ding, and Luo, 2013;

Ding et al., 2012), the impact of delays on fault detection (Dong et al., 2011; He, Wang, and Zhou, 2009; Zheng, Fang, and Wang, 2006), non-linear systems (Dong, Wang, and Gao, 2012a; Dong, Wang, and Gao, 2012b), complex networks (Gokas, 2000; Ding et al., 2013), basic fault diagnosis (Yin et al., 2012; Yin et al., 2013) or linear discrete time-varying systems (Zhong, Ding, and Ding, 2010).

Among various process supervision techniques, Fault Detection Diagnosis and Isolation (FDDI) is a significantly critical method for accomplishing a task to maintain a desirable performance in industrial processes, which commonly hold several kinds of faults. Most industries hope to improve their process performance through a higher level of FDDI capability. The basic functions of FDDI can be summed up into two parts, namely (1) monitoring the behavior of the process (variables) and (2) revealing the fault presence, its characteristics, and its causes. Thus, to maintain high process yield and throughput in industrial processes, it is necessary to adopt fast, accurate, and effective detection and diagnosis tools for process or equipment faults that may degrade the performance of the entire system. Many traditional FDDI techniques have been developed for checking the existence of a trend or pattern in the process or whether a certain process variable behaves normally or not. However, they might fail to produce several hidden characteristics of the process or fail to discover the faults in processes due to underlying process dynamics. The general implementation procedure of a conventional FDDI can be divided into four sequential steps as shown in Figure 2.1.

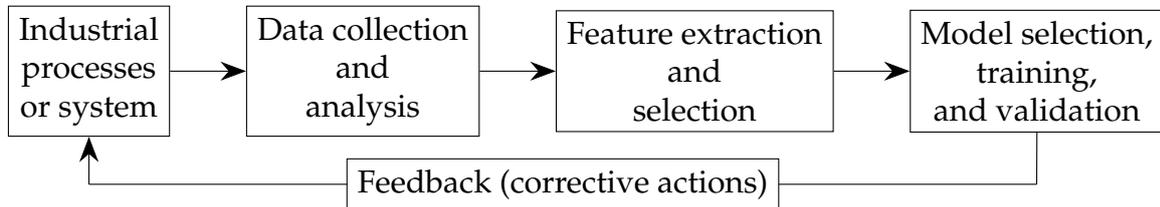


FIGURE 2.1: Implementation procedure of a conventional FDDI method

Fault diagnosis is a comprehensive task, as it has to determine fault type, fault size, location of a fault, time of the detected fault, and the behavior of faults through an appropriate assessment of the faults. Generally, fault diagnosis includes fault detection, isolation, identification, classification, and evaluation, but sometimes the combination of fault isolation and identification is called a fault diagnosis step (Park, Fan, and Hsu, 2020). Focusing on the data-driven FDDI approaches, they have received a lot of attention from diverse industries and have been widely applied in complex industrial process monitoring procedures (Yin et al., 2014; He, Wang, and Fan, 2018). The systematic categorization of various data-driven FDDI methods according to the system characteristics is given in Table 2.1.

Chapter 2. Research rationale

TABLE 2.1: Systematic categorization of data-driven FDDI methods

Dynamic System	Nonlinear System	Non-Gaussian System	Time-Varying/ Multimode System	Non-Stationary System
Hidden Markov model (HMM), Dynamic Principal Component analysis (DPCA), Discriminant partial least squares (DPPLS), Dynamic neural network (DNN), Filter methods	Kernel Principal Component analysis (KPCA), Kernel Independent Component analysis (KICA), Particle-filtering (PF), Artificial neural networks (ANNs), Kernel partial least squares (KPLS), Local Kernel Principal Component analysis (LKPCA)	Independent Component analysis (ICA), Gaussian mixture model (GMM), Kernel partial least squares (KPLS), Bayesian estimation	Recursive Principal Component analysis (RPCA), Recursive partial least squares (RPLS), Isolation estimation	Markov model (MM), Hidden semi-Markov model (HSM), Hidden Markov model (HMM), Monte Carlo simulation (MCS), Support vector machine (SVM)

Among various data-driven FDDI approaches, due to their simplicity and efficiency, Principal Component analysis-based (PCA) and partial least squares-based (PLS) FDDI methods are recognized as powerful tools to detect and diagnose process faults (Ding, 2014). In literature (Yin et al., 2014), one may find several data-driven methodologies, including design frameworks useful for monitoring and FDDI used in the process industry. For this, there are introduced process monitoring schemes to PCA- and PLS-based methods with successive necessary modifications needed for successful implementation of these schemes. Further, it is proposed to apply an integrated adaptive residual generation technique into the subspace-based process monitoring framework to resolve uncertainty problems. Wang et al., 2009 survey fault-tolerant control techniques as well as data-driven FDDI methods. In particular, they discuss the advances as well as general developments of data-driven FDDI and knowledge-based FDDI approaches widely used in many dynamic industrial processes. They provide a successful application example and presented future research directions, including many challenging issues in FDDI.

Yin et al., 2012 reviews several fundamental data-driven process monitoring and fault diagnosis (PM-FD) methods (PCA, PLS, ICA, and Fisher's discriminant analysis (FDA)). In this study, the developments and several characteristics of the mentioned data-driven methods are addressed, such as the original ideas, computational complexities, design, and algorithms in use. Qin, 2012 reviews the advanced developments of several data-driven PM-FD approaches and their applications. In this study, the Author discusses the use of a latent variable modeling approach and its extensions for fault detection. Also, he describes contribution plot approaches for fault diagnosis and identification and the contribution methods based on a reconstruction technique for fault identification, respectively. Furthermore, fault detectability and identifiability in fault diagnosis are discussed.

It is noted that the successful use of the above-given data-driven FDDI approaches depends on the identification of analytical models. There is always concern about noise and robustness, as well as the quality of historical data (Laouti, Sheibat-Othman, and Othman, 2011). It is mentioned that industrial systems are in the majority of cases nonlinear, thus there is a need to identify a corresponding nonlinear model. Nonlinear system identification is the art of determining a model of a nonlinear dynamical process by combining information obtained from data with that of physical insight or available *a priori* knowledge. There is a range of options in nonlinear system identification, and the choice of a particular method depends on the needs of the analysts, as well as the analyst's comfort with the fundamental ideas underlying a given tool. At the most basic level, the goal might be to merely identify how many states or modes are needed to construct a model of the system. With such information at hand, a more detailed system identification can start. At the more refined extreme is parametric system identification, for which the form of the differential equations of the system is known, but unknown parameters need to be identified. In between, these two extremes there lie techniques of non-parametric identification and nonlinear prediction, where the goal might range from revealing

Chapter 2. Research rationale

a nonlinear restoring force characteristic to modeling the dynamic behavior without determination of the differential equations.

Along with system identification, the model validation should come (Falkowski and Domanski, 2019b). It is a procedure that should be carried out before any attempt to explain and correct the differences, which have been observed between the test-derived model and its analytical counterpart. The validation is the process, which determines whether a given model is capable of describing the behavior of the system structure if all the individual model parameters are assigned to their correct values. A model may not be verified if it lacks certain features or degrees of freedom, which are present in the actual structure since, in this case, no amount of parameter correction can compensate for the errors embedded in the basic model. Models can be validated through the comparison of the responses of the model with the responses of the system to be identified. If possible, it is a good idea to use data separate from those used during the identification. Nonlinear system responses can be compared by characterizations of their responses, which is another research area. The nonlinear response can be very sensitive to the parameter estimation approach, modeling errors, and even the initial conditions. As such, validation based on responses should be performed with caution.

Considering the above, model identification as well as its subsequent validation, is a problematic and very subjective task (Falkowski and Domanski, 2019a). The analysis of large-scale industrial processes from the perspective of the control system seems to be a more universal approach. It is obvious that the main aim of a control system is to force a given set of process variables to behave in some desired and predefined way by either fulfilling some requirements in the time or frequency domain or achieving the best performances as expressed by an optimization index. Nevertheless, the scope of the control tasks varies widely. It may be to keep the process operation close to the nominal conditions. In other cases, the control purpose is to transfer the plant from one operating point to another or to track a given reference signal. In some other cases, the interest lies in how to obtain the best features of the plant achieving, for instance, the maximum production, minimum energy consumption or pollution, or minimum time in performing a given task. In a general way, the following control goals might be addressed (Astrom et al., 2001):

- process control (disturbance rejection),
- system setpoint tracking,
- the generation of the sequential procedures (for start-up or shut-down),
- system adaptation (changing some tunable parameters),
- process supervision (changing the operating conditions, structure, or components),
- system coordination (providing the operating points),

- process learning (extracting some knowledge from the experience),

and in that case, the most important is fault detection. It is stated that the control system of any process plays a critical role in achieving the desired objectives of the process, whether it is optimizing the process performance or achieving certain production targets. To perform these tasks, the control system utilizes various sensors to maintain the process. The information generated and used by the control system is not limited to the control variables. The system also generates valuable information about the process, such as disturbances, noises, process variability, and interactions between the different variables. This information is essential for the proper operation of the control system and can be used to identify and diagnose various issues that may arise in the process. Therefore, it is crucial to pay attention to this valuable information that the control system contains and to utilize it for different analyses. Consequently, the analysis performed from the control system perspective offers several advantages:

- control system exists and covers major process areas,
- each control loop contains major properties and the most important information about the process in a single variable,
- control error variable (an element of the loop) is representative for the loop, being already detrended, and thus trend-stationary (no trend),
- finally, it is known how to assess a single loop performance – an *atom* element of the multi-loop control system.

However, it is not as easy as it seems. Two issues relating to the analysis based on the control system emerge from the above considerations. The first thing is to measure control loop performance. In the literature, this issue is referred to as a Control Performance Assessment (CPA).

CPA adventure started with a simple univariate PID-based loop assessment. The first adequate report was proposed by (Åström, 1967) for a pulp and paper plant in 1967 using the benchmarking of process variable standard deviation. Control assessment solutions have evolved for more than 50 years in different directions, delivering to the industry mature approaches, measures, and procedures. There are many different representations of the industrial assessment process. Figure 2.2 shows a generalized diagram of the exemplary CPA industrial application process.

One may find a few methods' classifications in the literature and block tree diagrams visualizing functional similarities and differences. Figure 2.3 presents a graphical diagram of the generalized classification of the CPA techniques. The industrial perspective simplifies the picture (Domański, 2020). Simplicity is the main borderline, i.e., the scope of required *a priori* knowledge for the utilization of a selected approach. Methods that do not require specific knowledge can be simply evaluated by delivering a clear message. Again, the authors distinguish between

Chapter 2. Research rationale

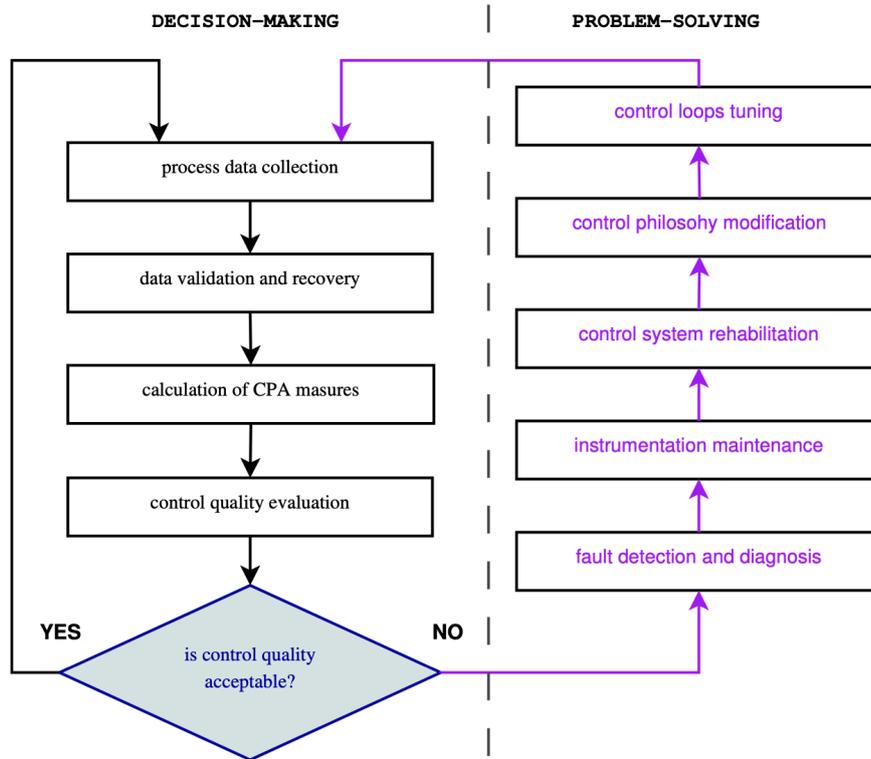


FIGURE 2.2: CPA industrial utilization process (Domański, 2020)

data-driven and model-based approaches. There are fundamental traps in the popular interpretation of these notions. First, each method uses data. Without data, there is no assessment, and actually, all methods are data-driven. Thereby, one might distinguish between model-free and model-based approaches. From that perspective, the majority of techniques are model-based only apart from some integral or time-based approaches.

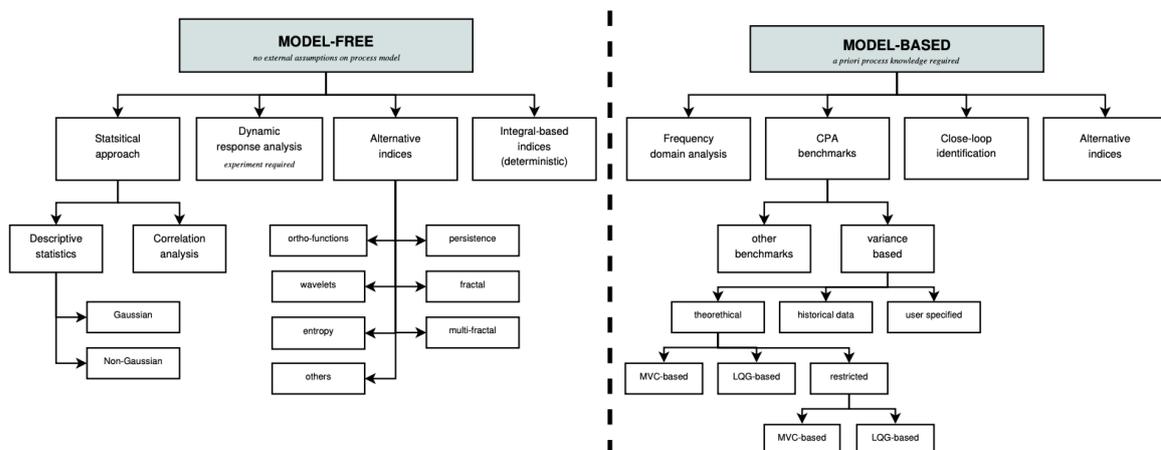


FIGURE 2.3: CPA techniques classification (Domański, 2020)

All the statistical approaches are model-based, as evaluated measures originate from some probabilistic density function (PDF), which is in fact a model. Thereby, the notion of the model has to be specified. A common understanding formulated by Miller and Starr, 1969 is that *the model is a representation of reality* (the real process). Consequently, the following classification is used:

- **Model-free** means that no process model is required;
- **Process model-based** approaches require performing the modeling of the controlled plant.

Therefore, model-free methods require only plant operational data, contrary to the process model-based approaches that always need some initial assumptions, for instance, model type, process delay, or structure. Moreover, the preferred methodology must be robust, i.e., it has to be independent of the existing loop characteristics noises, and statistical properties of the assessed variable. The goal is to measure internal control quality, not affected by any noise, disturbance, or possible plant influence of an unknown origin.

Present control performance assessment research encompasses various domains and applications to control engineering. Different methods' categories have been investigated in Betlem and Roffel, 2003; Shardt et al., 2012; O'Neill, Li, and Williams, 2017; Domański, 2020. The classification in Table 2.2 includes a short summary addressing the above-discussed issues.

Apart from the above items, there exists a group of methods utilizing hybrid or fusion approaches:

1. mixed CPA measures using sensor combination (Khamseh et al., 2016) or the Exponentially Weighted Moving Averages (EWMA) evaluated for other indexes (Salsbury and Alcala, 2015);
2. graphic visualization and patten recognition methods (Howard and Cooper, 2010; Dziuba et al., 2018; Domański et al., 2019); and
3. case-specific business Key Performance Indicators (KPIs), e.g., number of alarms or human interventions, time in manual mode (Knierim-Dietz, Hanel, and Lehner, 2012), and many other currency-based units (Bauer et al., 2016).

On the other hand, the analysis may not only address the control loop performance but may try to explain why the control system is not working properly and what is the actual cause of the error. It is noted that for both cases CPA analysis requires error propagation between loops. Therefore, a root cause analysis should be performed. This type of research is generally known as a Causality Analysis.

The Causality Analysis is the field of experimental design and statistics pertaining to establish the cause and the effect. Typically it involves the establishment of four elements: correlation, a sequence in time (that is, causes must occur before their effect), a plausible physical or information-theoretical mechanism for an

Chapter 2. Research rationale

TABLE 2.2: Methods classification based on different applications

Methods that require plant experiment	metrics that employ setpoint step response	overshoot, undershoot, rise time, peak time, settling time, decay ratio, offset (steady-state error), and peak value (Spinner, Srinivasan, and Rengaswamy, 2014)
	indexes that necessitate disturbance step response	Idle Index (Hägglund, 1999), Area Index, Output Index (Violi, 2006), and R-index (Salsbury, 2005)
Model-based methods	variance methods	minimum variance and normalized Harris index (Harris, 1989), Control Performance Index (Grimble, 2002), and other variance benchmarking methods (Harris and Seppala, 2001)
	model-based measures	aggressive/oscillatory and sluggishness indexes (Salsbury, 2007)
	frequency methods	Bode, Nyquist, and Nichols charts, with phase and gain margins Shardt et al., 2012, Fourier transforms (Schlegel, Skarda, and Cech, 2013), sensitivity functions (Tepljakov, Petlenkov, and Belikov, 2012), disturbance ratio index (Alagoz et al., 2015), and singular spectrum analysis (Yuan, 2015)
	alternative methods	indexes using neural networks (Zhou and Wan, 2008) and support vector machines (Pillay and Govender, 2017)
Data-driven methods	integral time measures	Mean Square Error (MSE), Integral Absolute Error (IAE) (Shinsky, 1990), Integral Time Absolute Value (ITAE) (Zhao, Xie, and Tu, 2012), Integral of Square Time derivative of the Control input (ISTC) (Zheng, 2007), Total Squared Variation (TSV) (Yu and Wang, 2016), and Amplitude Index (AMP) (Spinner, Srinivasan, and Rengaswamy, 2014)
	correlation measures	oscillation detection index (Horch, 1999) and relative damping index (Howard and Cooper, 2010)
	statistical factors	utilizing different probabilistic distribution function (standard deviation, variance, skewness, kurtosis, scale, shape, etc.) (Choudhury, Shah, and Thornhill, 2004), variance band index (Li and O'Neill, 2015), and factors of other probabilistic distributions (Zhong, 2003; Domański, 2017; Domański et al., 2018)
	benchmarking methods	(Hadjiiski and Georgiev, 2005)
	alternative indexes	wavelets (Nesic et al., 1997), orthogonal Laguerre (Lynch and Dumont, 1996) and other functions (Jelali, 2013), Hurst exponent (Pillay and Govender, 2014), persistence measures (Domański, 2016; Domański, 2019), entropy (Zhang, Jiang, and Chen, 2015; Zhou et al., 2018; Zhang et al., 2019), multifractal approaches (Domański and Gintrowski, 2017), and fractional-order (Liu et al., 2018; Liu, Chen, and Domański, 2020)

observed effect to follow from a possible cause, and the elimination of the possibility for common and alternative ("special") causes. The detection and clarification of cause-effect relationships among variables, events, or objects have been the fundamental question in the majority of natural and social sciences over the history of human knowledge. It is studied in dozens of high-impact domains including education (Dehejia and Wahba, 1999; Heckerman, Meek, and Cooper, 2006; Hill, 2011; LaLonde, 1986), medical science (Mani and Cooper, 2000; al., 2013), economics (Imbens, 2004), epidemiology (Hernan, Brumback, and Robins, 2000; Hernan and Robins, 2020; Robins, Hernan, and Brumback, 2000), meteorology (Ebert-Uphoff and Deng, 2012), and environmental health (Li, Zaiane, and Osornio-Vargas, 2014). Limited by the amount of data, solid prior causal knowledge is necessary to learn the causality. Researchers perform studies on data collected through carefully designed experiments, where solid prior causal knowledge is of vital importance (Heckerman, Meek, and Cooper, 2006).

There is no universally accepted definition of causality. The causality can be understood in terms of a "flow" among processes and expressed in mathematical language and mathematically analyzed. Most of the earlier research literature attempts to discuss unique causes in deterministic situations, and two conditions are important for deterministic causation: (i) the necessity: if X occurs, then Y must occur, and (ii) the sufficiency: if Y occurs, then X must have occurred. However, deterministic formulation, albeit appealing and analytically tractable, is not in accordance with reality, as no real-life system is strictly deterministic (i.e. its outcomes cannot be predicted with complete certainty). Thus, it is more realistic if one modifies the earlier formulation in terms of likelihood (i.e. if X occurs, then the likelihood of Y occurring increases). The first definition of causality, which could be quantified and measured computationally, yet very general, is given by Wiener, 1956: "For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one". The introduction of the concept of causality into the experimental practice, namely into analyses of data observed in consecutive time instants, time series, is due to Clive W. J. Granger, the 2003 Nobel prize winner in the economy. In his Nobel lecture (Granger, 2003) the Author recalls the inspiration from Wiener's work and identifies two components of the statement about causality:

- the cause occurs before the effect; and
- the cause contains information about the effect that is unique and is in no other variable.

As Granger puts it, a consequence of these statements is that the causal variable can help to forecast the effect variable after other data has been first used (Granger, 2003). This restricted sense of causality, referred to as Granger Causality, GC thereafter, characterizes the extent, to which a process X is leading another process Y, and builds upon the notion of incremental predictability (see Section 3.1.2).

Chapter 2. Research rationale

In physics and nonlinear dynamics, a considerable interest recently emerged in studying the cooperative behavior of coupled complex systems (Pikovsky, Rosenblum, and Kurths, 2001; Boccaletti et al., 2002). A nonlinear extension of the Granger causality approach is proposed by Chen et al., 2004 using local linear predictors. An important class of nonlinear predictors is based on the so-called radial basis functions (Butte and Kohane, 2000), which have been used for nonlinear parametric extension of the Granger causality concept (Ancona, Marinazzo, and Stramaglia, 2004; Marinazzo, Pellicoro, and Stramaglia, 2006). A non-parametric method to measure the causal information transfer between systems is proposed by Schreiber, 2000. His Transfer Entropy (TE) is designed as a Kullback–Leibler distance of transition probabilities (see Section 3.1.4). This measure is in fact an information-theoretic functional of probability distribution functions. It is used in climatology (Mokhov and A., 2006; Verdes, 2005), in physiology (Verdes, 2005; Katura et al., 2006), in neurophysiology (Chavez, Martinerie, and Le Van Quyen, 2003), and also in the analysis of financial data (Marschinski and Kantz, 2002), but never widely utilized in industrial engineering applications.

It is still possible to divide the methods into those, which are based on the model of an object and the model-free concepts. Among the previously developed as well as new solutions regarding causality, it would seem that the current spectrum of approaches in this research area is sufficient to conduct analyses in various fields of science. However, returning to engineering issues, causality concerning interconnecting control loops in complex multi-loop systems, as well as their propagation is associated with the existence of uncertainties and risks resulting from the nature of the analyzed signals. Control error signals may exhibit unknown properties. They can be crucial in proper root cause analysis and correct interpretation of the results. The most common, affecting the quality of the analysis, are:

- noises with the unknown stochastic process behind the data,
- outlying observations, and
- process data oscillations.

There are many intriguing questions that raise about this topic:

- How can we detect these phenomena?
- How do they affect the causality analysis?
- How should we circumvent or exclude them?
- How should we use them and what information do they bring into the picture?

These underlying questions accompanied by some of the minor issues are answered in this research. With the condition that the approach to the analysis of

large-scale industrial plants should be as general and universal as possible, the author focuses on the model-free Transfer Entropy method, which belongs to the group of data-driven methods. By avoiding the modeling issue, it is possible to reduce unnecessary degrees of freedom and potential base errors. The accuracy of a diagnosis depends not only on the diagnostic algorithm but also on the identification method used, making it crucial to consider these factors in the analysis. Relying solely on modeling can introduce a high complexity and lead to errors that may negatively impact the accuracy of the diagnosis. The following chapters present the characteristics of the most commonly used methodologies, at the same time justifying the choice of the TE method, through the analysis of simulation and real datasets.

Chapter 3

The research area review

Figure 3.1 illustrates a control system layout of a single-element control loop and presents its components, including a sensor, a controlled process, an actuator, and its controller. The main objective of the control system is to influence the behavior of the process in a desired way. The process produces an output that needs to be controlled, through the controller actions that aim at maintaining the measured process variable (CV) at a specified value defined by the setpoint (SP), and by taking into account (and rejecting) any disturbances impacting the process. The actuator, which contains the final control element, receives the controller output (OP) signal and influences the process in the desired manner.

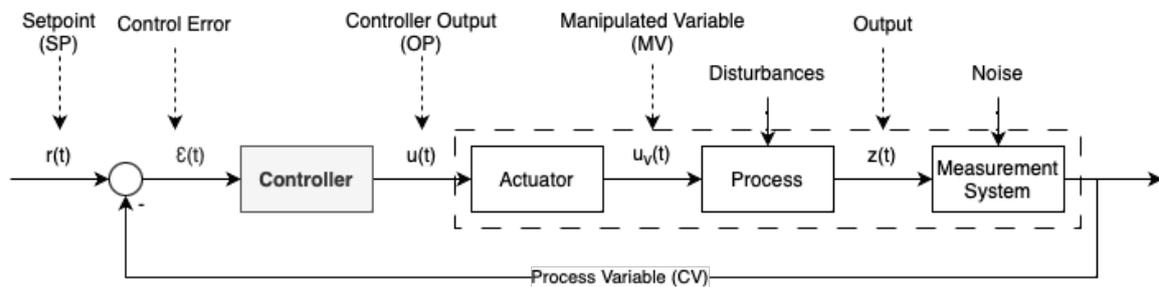


FIGURE 3.1: Component block diagram of a single element closed-loop system (Jelali, 2013)

To achieve the desired process control performance, it is crucial that all components of the control system function properly, and moreover their design is appropriate. However, ensuring the correct operation of each component is not a trivial task, even for a single-element control loop. From the perspective of a loop assessment, we need to identify and use specific information about its performance. We need to select the representative variables.

Considering a single-loop control system, the control error is considered the most valuable type of signal. This is because it is trend-stationary (Clements and Hendry, 2001). It means that it exhibits a stationary mean value over time, but has a non-zero covariance. There are several advantages of this property (Cochran, 2017):

1. Improved accuracy: Trend-stationary control error signals can help to provide a more accurate representation of the system's behavior by accounting for both the mean and variance of the error over time.
2. Better identification of system dynamics: By analyzing the trend or pattern of the control error signal, it is possible to gain insight into the underlying dynamics of the system, including the effects of external disturbances and changes in system parameters.
3. Enhanced control performance: Trend-stationary control error signals can be used to improve control performance by providing information that can be used to adjust control parameters and compensate for errors.
4. Increased stability assessment: Analyzing trend-stationary control error signals can help to identify instabilities in the system and guide the design of control systems that are more stable and robust.

Therefore, control error is easier to model and analyze mathematically, which is important in this application. The control error should fluctuate around zero in normal situations. If it is not there is a suspicion that the process is nonlinear or the control system is not designed properly (wrong operating point).

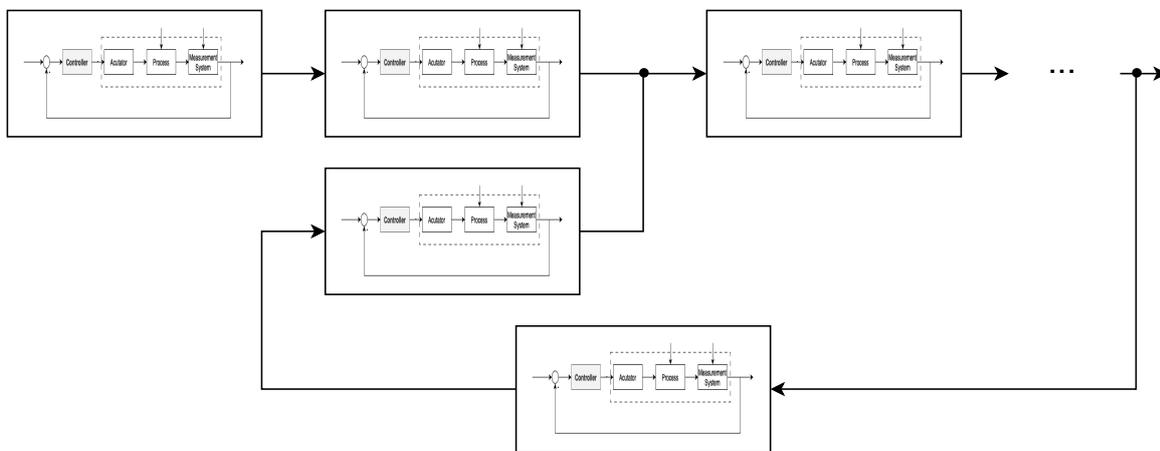


FIGURE 3.2: Component block diagram of a network of single-loop systems

The process industry typically comprises dozens of interconnected control loops (Figure 3.2). Such systems are called multi-loop. It is an enormous challenge to monitor and maintain top performance in such a situation. Sensors in large-scale industrial installations measure and record the values of process variables, generating multiple sets of respective time series (Lindner, Auret, and Bauer, 2017). This manifests in many challenges that must be faced during an analysis of the complex installation. This applies to both the method and the data. Therefore, both aspects must be addressed in the methodology.

Chapter 3. The research area review

Causality analysis should basically take into account the topology of the installation. It aims to reveal the relationships between respective control errors' time series and helps identify potentially causal connections between them. There are several techniques available for causality analysis used in control engineering, ranging from simple correlation methods to complex data-driven approaches. In this thesis, the author highlights the most popular and commonly used methods in the field. To select the appropriate approach it is required to start from the review of available techniques (Yang and Xiao, 2012). Section 3.1.1 describes the cross-correlation method, which involves the analysis of the similarity between two time series by shifting one of them and calculating the correlation coefficient. Section 3.1.2 presents the Granger causality method, which examines whether the past values of one-time series provide significant information for predicting the future values of another time series. Section 3.1.3 discusses the partial-directed coherence method, which measures the direct causal influence between two-time series, while taking into account the influence of other variables. However, the author places more emphasis on the Transfer Entropy approach, which is not widely applied in process analysis but serves as the basis for the analyses presented in this thesis. Section 3.1.4 delves into this method, which measures the amount of information that a time series provides about the future behavior of another time series, taking into account the influence of other variables. Discussion on the available solutions aims to present both their advantages and disadvantages concerning the Transfer Entropy approach that is chosen for the analysis.

Causality analysis is a critical tool for process analysis that aims to identify causal connections between different control errors' time series. However, the reliability and the accuracy of the results obtained using the Transfer Entropy method heavily depend on the accurate estimation of the Probabilistic Density Function (PDF) of the data. PDF can have a variety of non-Gaussian forms, and as such, understanding their impact on the results is critical. Unfortunately, research on alternative distributions in a control engineering context is limited, which limits our understanding of how different PDFs may affect the accuracy of the Transfer Entropy estimates. It is also important to note that fitting the distribution to the data is necessary, but not sufficient to guarantee optimal method performance. The fitting itself can be done using different approaches, like:

- the maximum likelihood (Coles and Dixon, 1999),
- the method of moments (“Method of Moments” 2008),
- the Goodness-of-Fit (D’Agostino and Stephens, 1986),
- the Q-Q plot method (Wilk and Gnanadesikan, 1968),
- the Least Squares estimation (Mert Kantar, 2015), or
- the L-moment Ratio Diagram (Hosking, 1990).

While this step helps to minimize the impact of data irregularities, it does not necessarily ensure that the results obtained are accurate. Therefore, it is important to carefully evaluate the quality of the estimates obtained and to take into account the limitations and assumptions of the Transfer Entropy method, which further uses these PDFs. Improving our understanding of the PDF and its impact on the TE estimates is essential across this application. An overview of the considered Probability Density Functions is included in Section 3.2.

After taking into account the importance of accurate estimation of the Probability Distribution Function in the Transfer Entropy approach, the next step is to consider the characteristics of the data itself. Large-scale industrial installations are often subject to various sources of disturbances, which can affect the behavior of the data. Contrary to initial assumptions, it cannot always be assumed that the signals do not exhibit any trend. When the setpoint changes, the control loop responds by adjusting the inputs to the system accordingly. This adjustment can result in significant changes in the behavior of the control errors, including the appearance of trends. In industrial applications, this issue is known and very popular. Ignoring it may lead to inaccurate conclusions and erroneous decisions. The methods of removing the trend, which are used in this task are described in Section 3.4.

Consistent with the claim that process data contains a lot of useful information, the presence of noises and oscillations in control errors can provide valuable insights into the behavior of large-scale multi-loop control systems. Although they are often perceived as unfavorable and undesirable, these signals can be used to gain a better understanding of the underlying processes. It is impossible to clearly say what is better to analyze and which type of signal better carries important information – the oscillation or the noise. Fortunately, there are available data processing techniques that can help to separate these effects from the raw data. One such technique is signal decomposition, which involves breaking down a complex signal into its constituent parts. This can help to identify noises and oscillations that may be difficult to discern otherwise. There are two main approaches to the signal decomposition: time-domain and frequency-domain decomposition. Time-domain decomposition involves separating the signal into different time intervals or segments, while frequency-domain decomposition involves breaking down the signal into its constituent frequencies. Both approaches have their advantages and limitations, and the choice of method depends on the specific application and the characteristics of the signal under consideration. Regardless of the approach used, signal decomposition is a powerful tool. Section 3.6 describes two considered approaches, which are used to solve this issue.

Poor tuning of control loops or actuator problems can lead to unstable or oscillatory behavior in the control system, which in turn can produce outliers in the collected process data. These outlying observations or anomalies can be caused by a variety of factors, including:

- data errors,

Chapter 3. The research area review

- measurement errors,
- uncoupled disturbances,
- human impact, etc.

The detection, labeling, and processing (eventual removal) of outliers are crucial for accurate analysis of the real-scale multi-loop control systems. Outliers can significantly bias statistical measures such as the mean and standard deviation, leading to incorrect conclusions about the behavior of the system. The most popular and proven methods related to the discussed issue are z-scores (also known as standard scores) or 3σ . The z-score is calculated by subtracting the mean of the population from the individual data point and then dividing the difference by the standard deviation of the population. This transformation allows the data to be expressed in terms of standard units, making it easier to compare data points from different populations or distributions. Z-scores are commonly used in statistical analysis, particularly in hypothesis testing, where they can be used to determine the probability that a given observation is due to chance, but also to identify outliers (Prosper, 2015). In the case of 3σ , it refers to three standard deviations above or below the mean. It is often used as a threshold for detection of outliers or unusual values in a data set. Specifically, a value that is more than three standard deviations away from the mean is considered an outlier or a point that is highly unlikely to have occurred by chance (Hacking, 2001).

Another common statistical method to identify outliers is to use Grubbs' test or Dixon's test, which compares the magnitude of the outlier to the variability of the data. If the outlier is deemed statistically significant, it can be detected, labeled, and eventually removed from the data set. Another approach for outlier detection is to use machine learning algorithms, such as clustering or anomaly detection techniques (Gupta et al., 2014; Mehrotra, Mohan, and Huang, 2017). These methods can be especially useful when dealing with large and complex data sets. Once outliers have been identified, they can be replaced with the mean or median value of the surrounding data points, or using interpolation techniques to fill in the missing values. In this case, three methods are proposed: extreme studentized deviate (ESD) (Section 3.5.1), interquartile range method (IQR) (Section 3.5.2), and Hampel filter (Section 3.5.3).

Chapter 4 lays the groundwork for the research presented in the following chapters by describing the simulation system and the initial analysis of the generated data, as well as the data obtained from a large-scale industrial installation. A simulation system is an essential tool for hypothesis testing, algorithms evaluation, or gaining a deeper understanding of complex systems. The chapter describes the system in detail, including the modeling of the process and the generation of data. The initial analysis of the generated data and the data obtained from the industrial installation provides valuable insights into the behavior of the system and highlights the challenges and limitations of the data. This analysis serves as a baseline for the

subsequent research and provides a context for the results presented in the following chapters. Chapter 5, Chapter 6, Chapter 7, Chapter 8, and Chapter 9 present the results of the research using the methods, algorithms, and approaches described in the thesis. These paragraphs develop the issues previously described in Chapter 2, Chapter 3, and Chapter 4, and provide a comprehensive analysis of the identification of causal relationships.

Finally, Chapter 10 contains a summary, supported by a case study using independent real data, while Chapter 11 presents concluding remarks that result from the conducted research and analyses.

3.1 Topology capturing methods

In the literature, the causality analysis of complex control systems uses the following approaches Cross-correlation (Cc), Granger Causality (GC), or Partial Directed Coherence (PDC). Apart from those mentioned, the Transfer Entropy (TE) approach is characterized as the basic method of further research.

3.1.1 Cross-correlation

The cross-correlation approach adjusts the time series of $x(t)$ and $y(t)$ so that there is a lag between them. The correlation coefficient between the adjusted time series is calculated, and the procedure is repeated for a range of assumed lags. The correlation $F_{x \rightarrow y}^{CC}$ for an assumed lag τ is calculated as shown in Eq. 3.1

$$F_{x \rightarrow y}^{CC} = \frac{1}{N - \tau} \sum_{i=1}^{N-\tau} \frac{(x_i(t) - \mu_x)(y_{i+\tau}(t) - \mu_y)}{\sigma_x \sigma_y}. \quad (3.1)$$

The N represents the number of samples and μ and σ represent the mean and standard deviation of the time series, respectively. The maximum correlation found is assumed to be the actual correlation between the time series. The lag that yields this maximum correlation is treated as an estimate of the time delay between the variables.

The advantage of the Cross-correlation approach lies mostly in its simplicity and its low computational demand (Lindner, Auret, and Bauer, 2017). However, the Cc can only detect linear interaction between time series. In large-scale industrial processes, where nonlinear behavior prevails, it may give inaccurate or even opposite results (Bauer and Thornhill, 2008). Moreover, the trend in a time series is ignored during the cross-correlation analysis. The data must be stationary, and the trend introduces non-stationarity (Yang et al., 2014). It means that the correlation between $x(t)$ and $y(t)$ at a specific lag τ is calculated, and the values of the time series of $x(t)$ and $y(t)$ are considered as the repeated measurements of the same event, instead of being measurements in time. This assumption about independent process, considered as independent realizations, may be hardly met in practice. Real-process

3.1. Topology capturing methods

industrial data often exhibit long-range dependence reflected in their persistence. This fact is contradictory with the independence assumption, and thus the Cc applicability is limited. Concluding, there is a high risk of an incorrect analysis of the real processes and a wrong interpretation of the results.

3.1.2 Granger Causality

The basic idea of the Granger Causality can be traced back to (Wiener, 1956), who conceived the notion that, if the prediction of one time series could be improved by incorporating the knowledge of a second one, then the second series is said to have a causal influence on the first one (Chen, Bressler, and Ding, 2009). Later Granger's formalization of this prediction idea in the context of linear regression models (Granger, 1969), says that regression of a variable on lagged values of itself is compared with the regression augmented with lagged values of the other variable. If the augmentation is helpful for better regression, then one can conclude that this variable is Granger-caused by another variable. For a better understanding, let's consider two variables (as time series), $x(t)$ and $y(t)$, that are jointly stationary. $y(t)$ is modeled as an autoregressive (AR) (Eq. 3.2), referred as the restricted one

$$y(t) = \sum_{i=1}^q B_i y(t-i) + \hat{\epsilon}_y(t). \quad (3.2)$$

Only the past values of $y(t)$ are used to predict future values; q is the model order defining the time lag, B is the AR polynomial coefficients, and $\hat{\epsilon}_y(t)$ is the prediction error. The unrestricted model of $y(t)$, which uses past values of both $x(t)$ and $y(t)$, is shown in Eq. 3.3

$$y(t) = \sum_{i=1}^q [B_{yx,i} x(t-i) + B_{yy,i} y(t-i)] + \hat{\epsilon}_{y|x}(t). \quad (3.3)$$

B_{yx} and B_{yy} represent the model polynomial coefficients, q is the model order. The Granger Causality is quantified as in Eq. 3.4

$$F_{x \rightarrow y}^{GC} = \ln \frac{\sigma_{\hat{\epsilon}_y}^2}{\sigma_{\hat{\epsilon}_{y|x}}^2}. \quad (3.4)$$

When the prediction of $y(t)$ is not improved with $x(t)$, $\sigma_{\hat{\epsilon}_{y|x}}^2 = \sigma_{\hat{\epsilon}_y}^2$ and $F_{x \rightarrow y}^{GC} = 0$. When the prediction of $y(t)$ is improved by including $x(t)$, $\sigma_{\hat{\epsilon}_{y|x}}^2 < \sigma_{\hat{\epsilon}_y}^2$ and $F_{x \rightarrow y}^{GC} > 0$; $x(t)$ is then said to be the Granger-cause $y(t)$.

The Granger Causality is characterized by its simple implementation and high efficiency when investigating causal relationships. Since the regression is common, the concept is easy to understand. Despite this, the method has its limitations, especially according to real-scale applications. The main one is associated with linearity,

thus in non-linear situations, the use of the Granger Causality is limited (Bressler and Seth, 2010). It means that the approach requires linear stochastic and wide-sense stationary (WSS) time series (Granger and Newbold, 1974). The causality measure accuracy depends on the exact regression model structure and their proper identification. The use of regression models introduces limitations to their applicability. The first is the prohibition of correlation, i.e. noise cannot be correlated with the input. This forces the analysis of open control loops and thus excludes the analysis of closed loops. The second limitation is the assumption that we deal with Gaussian disturbances. In large-scale real installations, it is rare. Restrictions also apply to model identification, and above all to the assumption that the order of the model polynomial is known. The subjective selection of the order of the polynomial raises doubts about whether errors in the analysis result from the model error or from the selected method of capturing the process topology. There are some extensions of the basic Granger Causality concept, which describes more general forms, but anyway, it should be used with caution and only in justified applications.

3.1.3 Partial Directed Coherence

Partial Directed Coherence (PDC) belongs to a group of frequency domain methods. Causality is represented by the energy transfer between pairs of time series at each frequency (Landman et al., 2014). PDC has been developed to provide a frequency domain description of Granger Causality (Baccala and Sameshima, 2001) (Section 3.1.2). The evaluation of directional interactions in the frequency domain is especially useful for a process with oscillatory behavior (Jiang, Patwardhan, and Shah, 2009). Based on time series, the k -dimensional autoregressive model is presented in Eq. 3.5

$$\begin{bmatrix} x_1(t) \\ \vdots \\ x_k(t) \end{bmatrix} = \sum_{i=1}^q \mathbf{B}_i \begin{bmatrix} x_1(t-i) \\ \vdots \\ x_k(t-i) \end{bmatrix} + \begin{bmatrix} v_1(t) \\ \vdots \\ v_k(t) \end{bmatrix}. \quad (3.5)$$

where $\mathbf{X}(t) = [x_1(t), x_2(t), \dots, x_k(t)]$ is the vector of k process variables, $\mathbf{Y}(t) = [v_1(t), v_2(t), \dots, v_k(t)]$ is k dimensional vector of the multivariate manipulated variable (MV) noise terms, $\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_i$ are $k \times k$ matrices of the model coefficients and q is the model order. The frequency response is obtained by application of discrete Laplace transform (Z-transform), setting $z^{-1} = e^{-j\omega}$ (Eq. 3.6)

$$\mathbf{B}(\omega)\mathbf{X}(\omega) = \mathbf{Y}(\omega), \quad (3.6)$$

$$\mathbf{X}(\omega) = x_1(\omega)x_2(\omega) \dots x_k(\omega), \quad (3.7)$$

$$\mathbf{Y}(\omega) = v_1(\omega)v_2(\omega) \dots v_k(\omega). \quad (3.8)$$

3.1. Topology capturing methods

The $\mathbf{B}(\omega)$ is the transfer function of the multivariate autoregressive model (Eq. 3.9)

$$B_{mn}(\omega) = - \sum_{i=1}^q b_{mn}(i) e^{-j\omega i} \quad (3.9)$$

and the Σ is the noise covariance matrix of the model (Eq. 3.10)

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \dots & \sigma_{1k}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \dots & \sigma_{kk}^2 \end{bmatrix}, \quad (3.10)$$

where σ_{cov} refers to covariance and σ^2 to the variance of the noise terms. It is assumed that the noise terms are uncorrelated.

If $b_{mn}(i) = 0$ for all values of i , there is no causality from x_m to x_n . The Partial Directed Coherence from variable x_m to x_n is defined by Eq. 3.11

$$|\hat{\tau}_{mn}(\omega)| = \frac{|B_{mn}(\omega)|}{\sqrt{\sum_{m=1}^n |B_{mn}(\omega)|^2}}. \quad (3.11)$$

The PDC is a function of $B_{mn}(\omega)$ alone and does not depend on the noise covariance matrix. It reveals only the power of the direct interactions between each pair of variables. The Partial Directed Coherence is useful for oscillatory data but still, an autoregressive model is present. Similar to GC, the causality measure depends on the accuracy of the model. Slowness in inference and generation of the model is observed. Computations based on the model are sequential (one at a time), so there is limited parallelism. Moreover, autoregressive models may introduce artificial bias when the wrong order is imposed. The PDC also makes assumptions about the linearity of the underlying dynamics, which limits its accuracy and applicability in industrial cases. Therefore, the same concerns as for GC apply to PDC and the method should be used with caution, in situations that fulfill the specific method's scope of applicability.

3.1.4 Transfer Entropy

The Transfer Entropy (TE) is an information-theoretic interpretation of Wiener's causality definition. In practice, this is a measure of information transfer from one variable to another that measures the reduction of uncertainty while assuming predictability (Schreiber, 2000). The basic formula of the TE is given by Eq. 3.12

$$T_{x \rightarrow y} = \sum_{y_{i+h}, \mathbf{Y}_i, \mathbf{X}_j} p(y_{i+h}, \mathbf{Y}_i, \mathbf{X}_j) \cdot \frac{p(y_{i+h} | \mathbf{Y}_i, \mathbf{X}_j)}{p(y_{i+h} | \mathbf{Y}_i)}, \quad (3.12)$$

where p means the complete or conditional Probabilistic Density Function (PDF), $\mathbf{Y}_i = [y_i, y_{i-l}, \dots, y_{i-(k-1)l}]$, $\mathbf{X}_j = [x_j, x_{j-l}, \dots, x_{j-(k-1)l}]$ are considered variables, l is

a sampling interval, and h is a prediction horizon. Thereby, it is the difference between information about a future observation of x obtained from the simultaneous observation of past values of both x and y . The information about the future of x is obtained only from the past values of x . It gives a good sense of the causality information without the need for any *a priori* process information, like a delay (Yang and Xiao, 2012). The phenomenon of bi-directional entropy in flow is highly probable. That is why a measure described as $T_{x \rightarrow y} = T_{x|y} - T_{y|x}$ is decisive due to quantity and direction, which is causality.

The practical implementation of the Transfer Entropy approach between a pair of variables according to Eq. 3.12 requires its simplification to the form presented in Eq. 3.13

$$T_{x \rightarrow y} = \sum_{y_i, y_{i-\tau_y}, x_{i-\tau_x}} p(y_i, y_{i-\tau_y}, x_{i-\tau_x}) \cdot \frac{p(y_i, y_{i-\tau_y}, x_{i-\tau_x}) p(y_{i-\tau_y})}{p(y_{i-\tau_y}, x_{i-\tau_x}) p(y_i, y_{i-\tau_y})}, \quad (3.13)$$

where p means the conditional PDF, τ_x and τ_y are the time delays in x and y , respectively. In the original formalization, the density function is based on Gaussian kernel density estimation. If the time series length is short, the τ_x is set to 1 under the assumption that the maximum auto-transfer of information occurs from the data point immediately before the target value in y .

In the classical formulation, Transfer Entropy between two variables (i.e. x and y) is calculated by function according to the methodology (Schreiber, 2000) presented in the form of

$$T_{x \rightarrow y} = \text{transferEntropy}(x, y, \tau_x, \tau_y, N_p, C_{thumb}), \quad (3.14)$$

where x and y are source time series as 1-D vectors, τ_x and τ_y are time delays of x and y respectively, N_p is a number of equally spaced points along each dimension where probabilities are estimated and C_{thumb} is the coefficient that adjusts the rule of thumb (Bauer et al., 2007). The Gaussian distribution is used for the probability density function in the original algorithm formulation. The function result is $T_{x \rightarrow y}$, which is the measure of information transfer between two given variables x and y .

Transfer Entropy naturally incorporates directional and dynamical information, because it is inherently asymmetric and based on transition probabilities (Vincente et al., 2011). The main advantage of such an information-theoretic functional causality detection formalization, in principle, does not assume any particular model for the interaction between variables. Thus, the sensitivity of Transfer Entropy to all order correlations becomes an advantage for exploratory analyses (Kayser, Sun, and Desposito, 2009) over model-based Cross-correlation, Granger Causality, or Partial Directed Coherence approaches. On the contrary, the TE method is a model-free approach. This is particularly relevant when the detection of some unknown non-linear interactions is required. In such areas, it is expected that it should give more accurate and robust results.

3.2. Review of parametric statistical models

On the other hand, the TE is restricted by a stationary formulation i.e.: the dynamic properties of the analyzed process do not change over the set of data used (Duan, 2014). The assumption that the time series is stationary does not hold and the noise (which may be nonstationary) is often greater than expected. The approach uses Probability Distribution Function formulation thus it is highly dependent on its accurate estimation. PDF can have any non-Gaussian form. Not many studies have ever been carried out for other distributions, and consequently, the impact of its change on the quality of the results obtained (Jafari-Mamaghani and Tyrcha, 2014). Fitting the distribution to the data does not preclude improving the performance of the method itself. Unfortunately, the resulting computational effort cannot be neglected. Calculations also involve numerous parameters e.g. prediction horizon h or the embedding dimensions.

Nevertheless, the Transfer Entropy is still a valuable method and offers great potential in applications in the analysis of real-scale multi-loop control systems. Using each of the methods described below, a causality diagram representing a given process can be determined; nodes represent measure variables (in this case, control loops error outputs), solid lines represent direct causal relationships between variables, and dashed lines represent indirect relationships (see Figure 3.3).

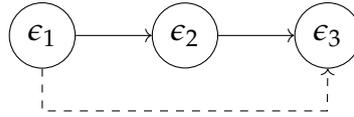


FIGURE 3.3: The sample layout of causality diagram

3.2 Review of parametric statistical models

The examination of various functions will commence with the standard Gaussian distribution, followed by the tailed functions of Laplace, α -stable family that includes the Cauchy function and the t Location-scale (by analogy with the others). Gaussian normal distribution description is extended with the robust moments' estimators. The choice is made since the mentioned models are well known, commonly used, and sufficient for most analyses using statistical approaches, including both simulation and real data.

3.2.1 Gaussian normal distribution

The normal probability distribution function is represented by a function of x and requires two parameters: the mean, denoted as μ , and the standard deviation, represented by σ (3.15). This function is symmetrical, and the mean serves as the offset

coefficient while the standard deviation determines the scale

$$F_{\mu,\sigma}^{Gauss}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (3.15)$$

Both function coefficients, i.e. μ and σ exist analytically. The Eq. (3.16) presents the results in a discrete-time case (x_1, \dots, x_{N_p}) , where N_p is a number of data points

$$mean(x) = \mu = \frac{1}{N_p} \sum_{i=1}^{N_p} x_i, \quad (3.16)$$

$$std(x) = \sigma = \sqrt{\frac{\sum_{i=1}^{N_p} (x_i - \mu)^2}{N_p - 1}}. \quad (3.17)$$

In Figure 3.4 there is an example of the Gaussian normal distribution as a function of location mean and standard deviation.

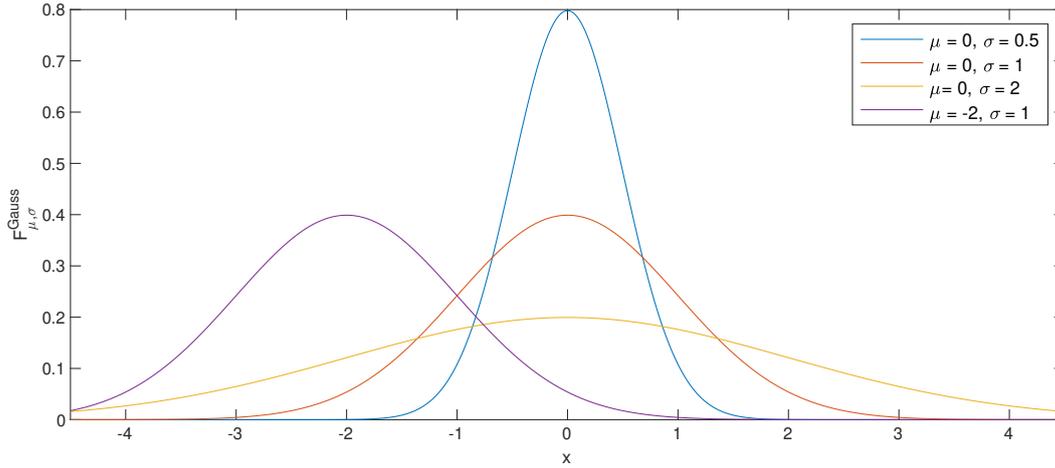


FIGURE 3.4: An illustrative representation of the Gaussian normal distribution as a function of mean and standard deviation

Changing the mean parameter causes the entire distribution to shift horizontally along the x -axis. If μ increases, the peak of the distribution shifts to the right, and if μ decreases, it shifts to the left. The mean represents the expected value of the distribution. Altering the standard deviation impacts the spread or dispersion of the data. When σ increases, the distribution becomes wider, and data points are more spread out from the mean. Conversely, if σ decreases, the distribution narrows, and data points cluster closer to the mean. Changes in μ and σ also influence the skewness and the kurtosis of the distribution.

The skewness is a statistical measure that quantifies the asymmetry or lack of symmetry in the probability distribution of a dataset. In essence, it assesses the

3.2. Review of parametric statistical models

extent to which the data deviates from a perfectly symmetrical, bell-shaped, normal distribution. The formula of skewness is given by Eq. 3.18

$$skewness(x) = \beta = \frac{1}{N_p} \sum_{i=1}^{N_p} \frac{(x_i - \mu)^3}{\sigma^3}, \quad (3.18)$$

where μ is mean, σ is standard deviation, and N_p is the number of data points.

It is computed by examining the third standardized moment of a dataset. The standardized moment is calculated by dividing the third moment (the average cubed deviation from the mean) by the cube of the standard deviation. The resulting value can be positive, negative, or zero, each conveying specific information about the distribution.

The kurtosis characterizes the peakedness or flatness of the probability distribution of a dataset in comparison to a normal distribution. It provides insights into the shape of the distribution's tails and the presence of outliers. The formula of the kurtosis is given by Eq. 3.19:

$$kurtosis(x) = \eta = \frac{1}{N_p} \sum_{i=1}^{N_p} \frac{(x_i - \mu)^4}{\sigma^4}, \quad (3.19)$$

where μ is mean, σ is standard deviation, and N_p is the number of data points.

Kurtosis is computed by examining the fourth standardized moment of a dataset, which is obtained by dividing the fourth moment (the average fourth power of deviations from the mean) by the fourth power of the standard deviation.

One important property of Gaussian distribution is that it is sensitive to outliers, which means that the presence of even a few extreme values can significantly affect the mean and variance of the distribution. On the other hand, a method that has some similarities to the Gaussian statistical model, but is much more robust, is Huber's logistic estimator.

3.2.2 Robust statistics – Huber's logistic estimator

The existence of outliers in data implies fat tails in their distributions (Domański, 2020). This feature biases the standard estimation of normal moments. There are many other estimators for the basic moments: i.e. the mean and standard deviation. The median value is considered as a simple and robust alternative to the mean value

$$median(x) = \tilde{\mu} = \begin{cases} \frac{x_{\frac{N_p}{2}:N_p} + x_{\frac{N_p}{2}+1:N_p}}{2} & \text{if } N_p \text{ is even,} \\ x_{\frac{N_p+1}{2}:N_p} & \text{if } N_p \text{ is odd,} \end{cases} \quad (3.20)$$

where $x_{1:N_p} \leq x_{2:N_p} \leq \dots \leq x_{N_p:N_p}$ are the ordered observations.

There are quite many robust scale estimators (Rousseeuw and Leroy, 1987). In the considered analysis author uses the M-estimator with Huber ψ -function (Croux and Dehon, 2014).

Scale M-estimator is obtained by solving Eq. (3.21)

$$\frac{1}{N_p} \sum_{i=1}^{N_p} \rho \left(\frac{x_i - \tilde{\mu}}{\gamma} \right) = \kappa, \quad (3.21)$$

where $0 < \kappa < \rho(\infty)$, $\rho(\cdot)$ is even, differentiable, and non-decreasing on the positive numbers for the loss function, γ is a scale estimator and $\tilde{\mu}$ is a preliminary shift factor (median). While the logistic ψ function (3.22) is taken as $\rho(\cdot)$ we obtain the logistic ψ scale estimator

$$\psi_{log}(x) = \frac{e^x - 1}{e^x + 1}. \quad (3.22)$$

3.2.3 The family of α -stable distributions

The α -stable distribution lacks a closed-form probability density function and is instead defined through its characteristic function as given in Eq. (3.23)

$$F_{\alpha,\beta,\delta,\gamma}^{stab}(x) = \exp \{ i\delta x - |\gamma x|^\alpha (1 - i\beta f(x)) \}, \quad (3.23)$$

where

$$f(x) = \begin{cases} \operatorname{sgn}(x) \tan\left(\frac{\pi\alpha}{2}\right) & \text{for } \alpha \neq 1 \\ -\operatorname{sgn}(x) \frac{2}{\pi} \ln|x| & \text{for } \alpha = 1 \end{cases}$$

$0 < \alpha \leq 2$ called the *stability index* or the characteristic exponent,

$|\beta| \leq 1$ is a *skewness factor*,

$\delta \in \mathbb{R}$ is a distribution *location* or *shift factor*,

$\gamma > 0$ is distribution *scale*.

The α -stable distribution function is described by four parameters: the shift δ , the scale γ and two shape factors: α and β . However, there are certain special cases, where the probability density function (PDF) has a closed form:

- $\alpha = 2$, it represents independent realizations, in particular, when $\alpha = 2$, $\beta = 0$, $\gamma = 1$, and $\delta = 0$, the exact equation for normal distribution is obtained,
- $\alpha = 1$ and $\beta = 0$, it represents the Cauchy distribution, which will be discussed in detail in the following paragraph, and
- $\alpha = 0.5$ and $\beta = \pm 1$, it represents the L'evy distribution, which is not included in the analysis.

3.2. Review of parametric statistical models

In Figure 3.5 there is an example of the α -stable distribution as a function of stability index, skewness parameter, location factor, and scale.

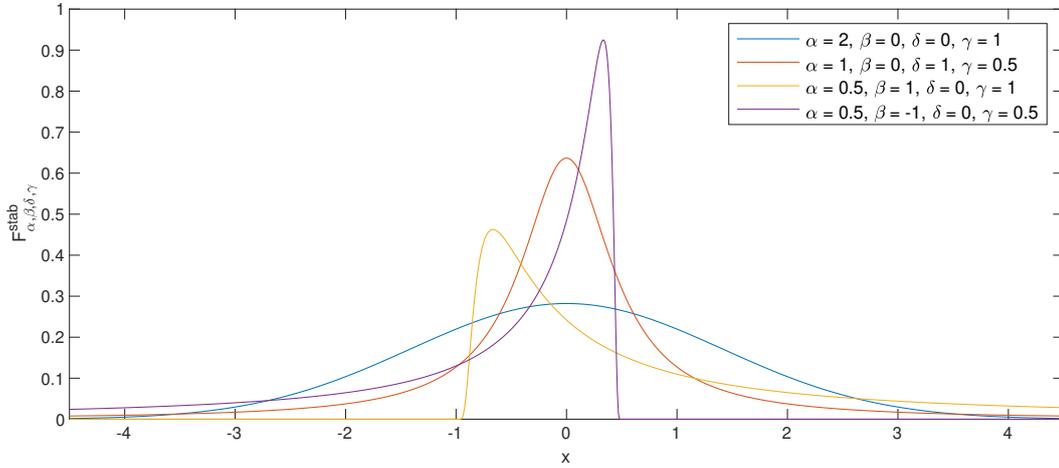


FIGURE 3.5: An illustrative representation of the α -stable distribution as a function of stability index, skewness, location factor, and scale

Cauchy probabilistic density function

Cauchy PDF is an example of the fat-tail distribution belonging to the family of stable distributions. The distribution is symmetrical with the following density function

$$F_{\delta, \gamma}^{\text{Cauchy}}(x) = \frac{1}{\pi\gamma} \left(\frac{\gamma^2}{(x - \delta)^2 + \gamma^2} \right), \quad (3.24)$$

where

$\delta \in \mathbb{R}$ is a distribution *shift* (offset) parameter,

$\gamma > 0$ is distribution *scale* factor.

In Figure 3.6 there is an example of the Cauchy distribution as a function of location parameter and scale factor.

The location parameter induces a horizontal shift of the graph along the x-axis, while the scale factor influences the shape of the curve, making it either more compact or more stretched. A smaller scale parameter leads to a taller and narrower curve.

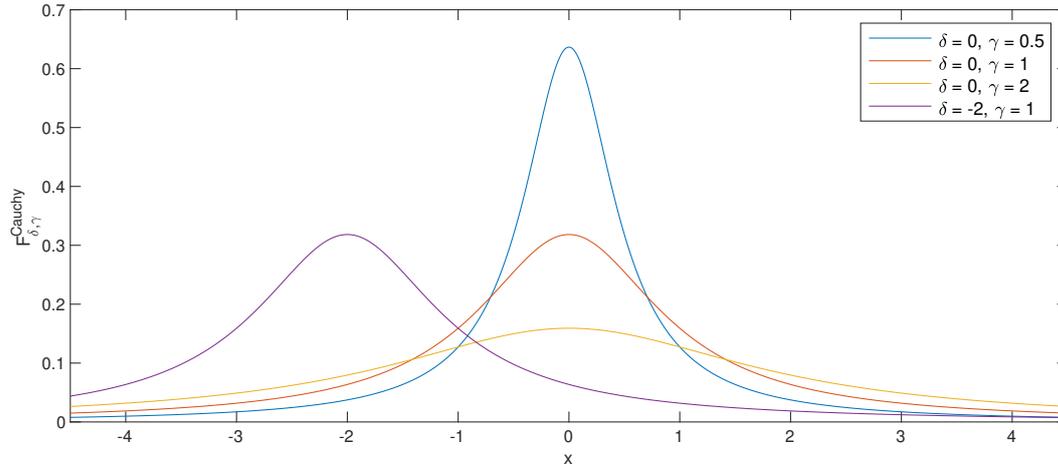


FIGURE 3.6: An illustrative representation of the Cauchy distribution as a function of location parameter and scale factor

Laplace double exponential distribution

The Laplace distribution, also known as the double exponential distribution, is constructed as the difference between two independent variables with identical exponential distributions. Its probability density function is expressed as:

$$F_{\delta, \gamma}^{Lap}(x) = \frac{1}{2\gamma} e^{-\frac{|x-\delta|}{\gamma}}, \quad (3.25)$$

where $\delta \in \mathbb{R}$ is a distribution *shift* (location) factor and $\gamma > 0$ is a scale parameter.

Figure 3.7 demonstrates an illustrative representation of the Laplace double exponential distribution, influenced by variations in the offset factor and scale parameter.

This particular distribution arises from the convolution of two exponential distributions, with one being positively skewed and the other negatively skewed. The offset factor is responsible for imparting a horizontal translation to the graph along the x-axis, whereas the scale parameter governs the curve's shape, determining whether it appears more compressed or elongated. A reduced scale parameter results in a taller and more slender curve. Notably, the Laplace double exponential distribution exhibits a sharper peak compared to both the normal and Cauchy distributions.

3.2. Review of parametric statistical models

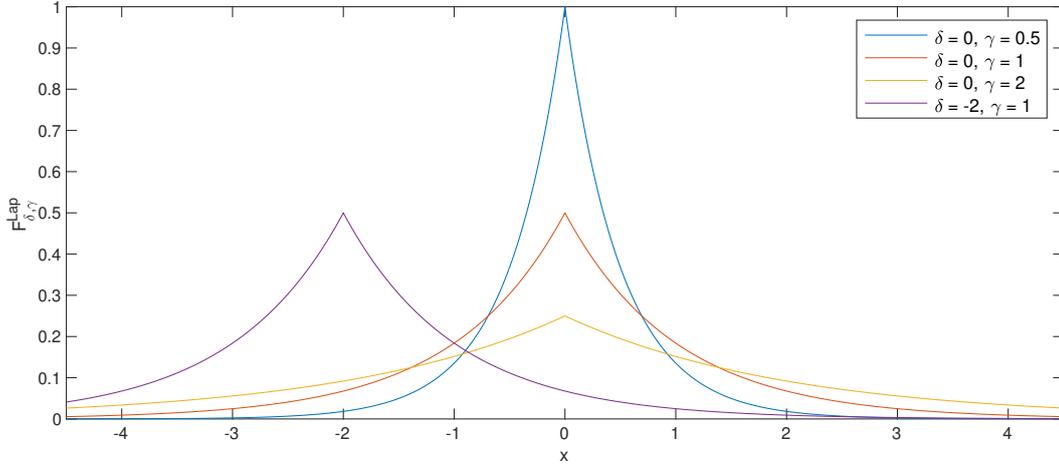


FIGURE 3.7: An illustrative representation of the Laplace double exponential distribution as a function of offset factor and scale parameter

t Location-scale distribution

The location-scale version of the t-distribution includes more degrees of freedom than the standard t-distribution. It is described by the following functions:

$$F_{\delta, \gamma, \nu}^{t-LS}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sigma \sqrt{\nu\pi}} \left(\frac{\nu + \left(\frac{x-\mu}{\sigma}\right)^2}{\nu} \right), \quad (3.26)$$

where

$\delta \in \mathbb{R}$ is a distribution *shift* (offset) factor,

$\gamma > 0$ distribution *scale*,

$\nu > 0$ function *shape*,

and $\Gamma(\cdot)$ denotes a Gamma function.

The mean of the t Location-scale distribution is defined by δ , and the variance by

$$\text{var} = \sigma^2 \frac{\nu}{\nu-2}. \quad (3.27)$$

The variance is only defined for values of $\nu > 2$. For other values of ν , the variance is undefined.

Figure 3.8 illustrates an example of the t Location-scale distribution, with considerations given to variations in the location factor, scale parameter, and shape parameter.

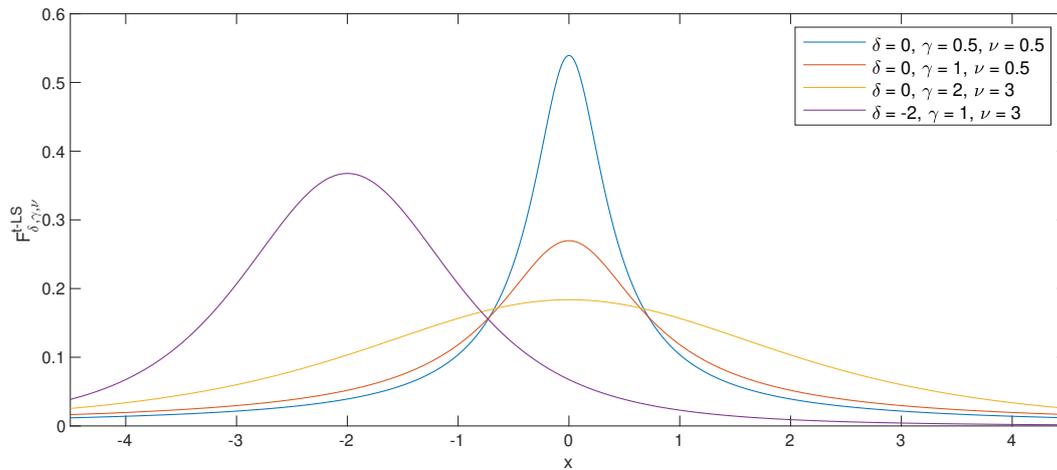


FIGURE 3.8: An illustrative representation of the t Location-scale distribution as a function of location factor, scale parameter, and shape parameter

The location factor determines the horizontal shift of the distribution along the x -axis. The scale parameter influences the spread or dispersion of the distribution. A smaller scale parameter results in a narrower, more peaked distribution, while a larger scale parameter leads to a broader, more spread-out distribution. The shape parameter is specific to the t distribution and is responsible for controlling the shape of the distribution's tails. It influences the thickness of the tails relative to the center of the distribution. A lower shape parameter leads to heavier, fatter tails, resembling a Cauchy distribution, while a higher shape parameter results in thinner tails, approaching the behavior of a normal distribution.

3.3 Review of non-parametric statistical models

Apart from the widely recognized and frequently employed statistical models, alternative non-parametric models can be encountered in the literature (Kraskov, Stögbauer, and Grassberger, 2004; Lizier, 2014). They are more flexible and can capture not obvious patterns and relationships that are not captured by the parametric statistical models. Moreover, non-parametric models may have better generalization performance, especially when dealing with complex and high-dimensional data. They can handle situations, where the relationship between variables is nonlinear or where the interactions are present, leading to more accurate predictions or estimations. Therefore, unusual approaches, such as Darbellay-Vajda, Fixed-Bins, and Kernel Density Estimation algorithms are characterized.

3.3. Review of non-parametric statistical models

3.3.1 Darbellay-Vajda algorithm

The Darbellay-Vajda (DV) algorithm is an adaptive histogram-generating process that is determined by partitioning the observation space into a finite number of the non-overlapping rectangular cells that are received by the recursive process (Darbellay and Vajda, 1999). This approach relies on an iterative partitioning of the cells on the observation space, based on a chi-square statistical test to ensure conditional independence of the proposed partitioned cells from the rest of the cells.

3.3.2 Fixed-Bins algorithm

The simplest estimation approach to obtain the PDFs is called the Fixed-Bins algorithm. It allocates data points to fixed, equally-spaced bins. To enhance robustness against outliers and sparse regions in the underlying distribution, there is combined fixed binning with ordinal sampling (ranking). In ordinal sampling, the two analyzed time series values are substituted with their ranks in those time series, similar to most non-parametric statistical tests (Lee et al., 2012). The ranks are integers ranging from the smallest to the largest value. The difficulty arises in selecting the optimal number of quantization levels Q for the analyzed pair of signals (i.e. x and y) in the Fixed-Bins algorithm due to the lack of a priori knowledge regarding coupling time lag.

To select the optimal quantization level Q for the analyzed pair of signals x and y , the constant time lag for signal y is set, and a significant information transfer is observed while increasing the time lag in x . This observation highlights the importance of appropriately adjusting the time lag to capture the underlying relationship between the signals. The example of a parameter Q selection for the fixed-bins algorithm is presented in Figure 3.9.

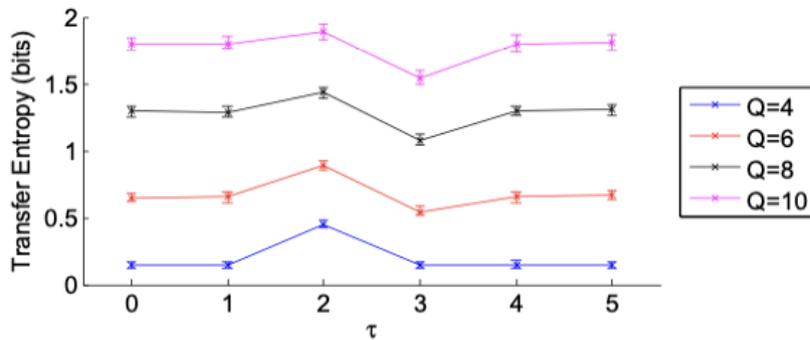


FIGURE 3.9: Parameter selection for the fixed-bins algorithm (taken from Lee et al., 2012)

The example in Figure 3.9 reveals that the highest Transfer Entropy values are obtained at a specific time lag, $\tau = 2$, indicating a notable information flow between the signals at this particular delay. Selection of the quantization level is based on the relatively highest Transfer Entropy coefficient. In this case, it will be $Q = 4$.

3.3.3 Kernel Density Estimation algorithm

Kernel Density Estimation (KDE) is utilized to produce a smoothed PDF estimation using the data samples, which stands in contrast to the histogram model which has sharp edges resulting from a uniform distribution within each bin (Gencaga, Knuth, and Rossow, 2015). In this method, a preselected distribution of values around each data sample is summed to obtain an overall, smoother PDF in the data range. This preselected distribution of values within a certain range is known as a "kernel" (Silverman, 1986). Some of the most commonly used kernels are Rectangular or Gaussian. Each kernel can be thought of as a window with a bandwidth or radius. Even if a Gaussian kernel is used, the resulting entropy estimation is more accurate compared to the classic histogram approach (Prichard and Theiler, 1995). The problem is that this method has a free parameter, λ , which is a multiplier for scaling in PDF based on Gaussian Kernel Density Estimation and must be selected to estimate the neighborhoods (Kugiumtzis, 2009; Prokopenko and Lizier, 2014).

The selection of this parameter is similar to choosing the fixed bins width in a histogram. The constant time lag for signal y is set, and a significant information transfer is observed while increasing the time lag in x . The example of a parameter λ selection for the KDE algorithm is presented in Figure 3.10.

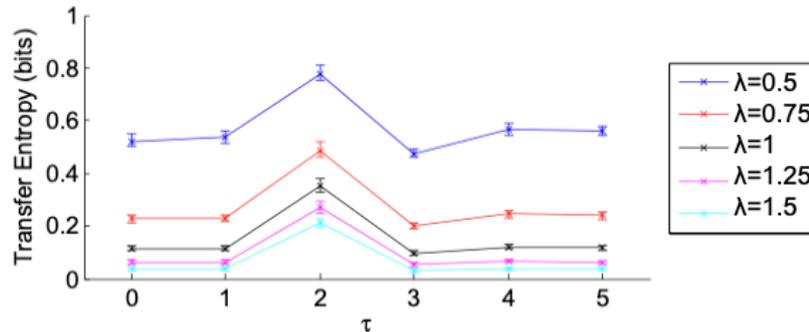


FIGURE 3.10: Parameter selection for the KDE algorithm (taken from Lee et al., 2012)

The observed significant information flow between the signals is specifically observed at a time lag of $\tau = 2$, which is supported by the highest Transfer Entropy values obtained at this particular delay. This finding justifies the choice of the parameter $\lambda = 1.5$, as it corresponds precisely to the quantization level that yields the relatively largest Transfer Entropy value.

3.4 The trend identification

The trend identification and its removal can be made by using the most common polynomial interpolation. It is a powerful mathematical tool for approximating functions and estimating values within a given range. Polynomial interpolation

3.5. Outliers detection

is typically performed using a process called Lagrange interpolation. This involves constructing a polynomial that passes through each of the data points, by finding a set of coefficients that satisfy certain conditions. The degree of the polynomial is determined by the number of data points being used for the interpolation.

Choosing the right i^{th} polynomial degree is a contentious issue. Objective evaluation of a selection of appropriate polynomial degrees is carried out by applying the concept used for a different task and presented in (Taleb, 2020). An increasingly complex trend is set, and then after its removal, the mean absolute deviation (MAD) of a signal is calculated (Falkowski, 2022). It is expressed as:

$$MAD(x) = \frac{1}{N_p} \sum_{i=1}^{N_p} |x_i - \mu|, \quad (3.28)$$

where μ is mean and N_p is the number of data points.

If MAD converges to a relatively small and constant value, then an order of the polynomial is determined. Regardless of the selected degree for the analyzed signal, its MAD may not change significantly. This is why polynomial interpolation is sensitive to noise in the data, and may not always produce accurate results. In some cases, setting the right polynomial order is impossible, and alternative interpolation methods, such as spline interpolation or radial basis function interpolation, may be more appropriate.

The spline interpolation, which is defined piece-wise by polynomials (Schumaker, 1981), has the ability to handle non-uniformly sampled data and can be used to extrapolate beyond the boundaries of the data. This is especially useful when dealing with noisy or incomplete data sets. This method works by constructing a series of piece-wise polynomial functions that smoothly connect the data points. These polynomials are selected in such a way that they minimize the curvature of the function between the data points, resulting in a smooth and continuous function. An advantage of spline interpolation is that it allows for the easy computation of derivatives and integrals of the interpolated function, which is useful in this application. Spline interpolation offers a powerful alternative that is flexible, robust, and can handle a wide range of data sets.

3.5 Outliers detection

An outlier is a strange occurrence commonly named as *"an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"* (Hawkins, 1980). This phenomenon has a significant effect on data analysis. They increase signal variance and reduce the power of statistical tests during performed analysis (Osborne and Overbay, 2004), destroy signal normality, and introduce fat tails (Taleb, 2020) and bias regression analysis (Rousseeuw and Leroy, 1987). Outliers come from erroneous observations or inherent data variability.

Outliers should be detected during data analysis to label them for further activities. In the current work, the detection of abnormal data behavior is used to "clean" data from the outliers before the core process of the root cause analysis. It is shown that outlier detection improves causality analysis. Three approaches are used: generalized extreme studentized deviate (ESD), interquartile range method (IQR), and Hampel filter. Their description is presented below.

3.5.1 Extreme studentized deviate

The generalized extreme studentized deviate test (Rosner, 1983) can be used to find one or more outliers in data originating from an approximately normal distribution. It assumes the upper limit of the outliers number. Knowing the upper limit of the outliers number, N_o , the test performs N_o separate tests (similar to Grubb's test): for one, for two, and so on up to N_o outliers. Generally, the extreme studentized deviate test is the sequentially applied Grubb's test. However, it performs appropriate tunings for critical values taking into account the number of tested outliers. Finally, it is robust to the significant masking effect. Detected outliers are replaced by threshold-level values.

3.5.2 InterQuartile Range method

The majority of data is not normal enough to be considered as drawn from a normal distribution. A possible statistic is the IQR approach (Whaley, 2005) in such a case. It is evaluated as a difference between upper 75th (Q_3) and lower 25th (Q_1) percentile of a dataset. IQR may be used to find outliers. They are considered as observations that fall below $LL = Q_1 - 1.5$ or above $HH = Q_3 + 1.5$ thresholds. Detected outliers are replaced by threshold-level values.

3.5.3 Hampel filter

Hampel filter (Pearson, 2005) as such is a sliding window version of the classical MDist (3σ) algorithm with a median estimator of location. The MDist uses the well-known property of the normal distribution that if x is distributed with $N(\mu, \sigma^2)$, then $z = (x - \mu)/\sigma$ is distributed with $N(0, 1)$. Thereby, one may use the Z-scores of the observations x_1, x_2, \dots, x_N ,

$$z_i = \frac{x_i - \mu}{\sigma} \quad (3.29)$$

as an algorithm to label outliers.

The common rule labels MDist that exceed 3 in absolute value as outliers. Generally, the Hampel filter is used to detect local outliers. Being a standard median filter with a symmetric moving window it uses one parameter: the window width. Detected outliers are replaced by the median of neighbors within the window.

3.6 Data decomposition methods

During the analysis, two decomposition approaches are considered: Ensemble Empirical Mode Decomposition (EEMD) and its extension Median Ensemble Empirical Mode Decomposition (MEEMD).

3.6.1 Ensemble Empirical Mode Decomposition

Since there is usually noise and signal intermittency in real-world data, it causes mode mixing and mode splitting (MS) using EMD method. To avoid this issue the Ensemble Empirical Mode Decomposition (EEMD) (Wu and Huang, 2009) is proposed. EEMD belongs to a class of noise-assisted EMD methods that are the most powerful tools for performing time-frequency analysis (Mandic et al., 2013; Rehman and Mandic, 2009). Those are aimed at alleviating mode mixing caused by noise and signal intermittency and can be successfully used in the noise and oscillation identification process.

3.6.2 Median Ensemble Empirical Mode Decomposition

Median Ensemble Empirical Mode Decomposition (MEEMD) is a variation of the EEMD algorithm that uses the median operator instead of the mean operator to ensemble noisy Intrinsic Mode Function (IMF) trials (Lang et al., 2020). The use of this algorithm is a practical extension of the classic EMD and a justified choice with real-world applications. The EMD method was developed so that data can be examined in an adaptive time-frequency–amplitude space for nonlinear and non-stationary signals (Mandic et al., 2013). It decomposes the input signal into a few Intrinsic Mode Functions and a residue. The given equation will be as follows:

$$I(t) = \sum_{i=1}^{N_{IMF}} IMF_i(t) + Res_{N_{IMF}}(t), \quad (3.30)$$

where $I(t)$ is the multi-component signal. $IMF_i(t)$ is the N_{IMF}^{th} intrinsic mode function, and $Res_{N_{IMF}}(t)$ represents the residue corresponding to N_{IMF} intrinsic modes. The proposed median EEMD (MEEMD) algorithm, defines the median operator as:

$$median = \begin{cases} IMF(t)[(N_{IMF} + 1)/2] & \text{– when } N_{IMF} \text{ is odd,} \\ \frac{IMF(t)[(N_{IMF}/2)] + IMF(t)[(N_{IMF}/2)+1]}{2} & \text{– when } N_{IMF} \text{ is even,} \end{cases} \quad (3.31)$$

where $IMF(t)$ denotes the ordered IMF list at time instant t , which is obtained from N_{IMF} independent noise realizations. Consider a real-valued time series $x(t)$ and a predefined noise amplitude v , the used MEEMD is outlined in Algorithm 1 (Falkowski, Domański, and Pawłuszewicz, 2022).

Algorithm 1 Algorithm of MEEMD

1. Generate the ensemble $S_i(t) = x(t) + v\omega_i(t)$ for $i = 1, \dots, N_{IMF}$, where $\omega_i(t) \sim N(0, 1)$;
 2. Decompose every member of $S_i(t)$ into N_{IMF_j} IMFs using the standard EMD, to yield the set $(d_j^i(t))_{i=1}^{N_{IMF_j}}$;
 3. Assemble same-index IMFs across the ensemble using the median operator to obtain the final IMFs within MEEMD; for instance, the j^{th} IMF is computed as $d_j(t) = \text{median}(d_j^1(t), d_j^2(t), \dots, d_j^{N_{IMF}}(t))$.
-

Chapter 4

Description of the systems

In this Chapter, a detailed overview of the simulation system is provided, which is used to conduct the research. Next, the real industrial system is characterized, called a case study example, to demonstrate the effectiveness of the analytical procedure proposed and verified in Chapter 10.

4.1 Description of the simulation system

To perform an analysis of simulation data, a sample multi-loop simulation model is created in MATLAB Simulink[®] environment. The model scheme is presented in Figure 4.1. This model includes different types of processes with their transfer functions taken from Åström and Hägglund, 2000 (PID control benchmarks) and from Bequette, 2002 (boiler drum model). The system combines feedback PID control SISO loops, cascaded structures, and feedforward filters. The model consists of five loops controlled by four PID controllers (R_1 , R_2 , R_3 , and R_4) and one PI (R_5) controller. The control errors from the loops, denoted in the scheme respectively as ϵ_1 , ϵ_2 , ϵ_3 , ϵ_4 , and ϵ_5 , respectively, are the basis for further evaluation. The model incorporates the possibility of applying noises. In this case, two variants are given. First, the simulation data can be distorted by the Gaussian noise, which reflects the measurement noise (represented by variable σ_{G_i}), and second, the process may be distorted by the Cauchy (represented by variable γ_{C_i}) disturbances as well. Such fat-tailed noise represents industrial disturbances as shown in Domański, 2015. A sinusoidal signal can also be added to the Cauchy disturbance to simulate potential loop oscillations of known frequency to assess their propagation through the multi-loop structure.

Simulated process transfer functions are in the form of the following linear models:

$$\begin{aligned} G_1(s) &= \frac{1}{0.15s + 1}, & G_2(s) &= \frac{0.25(-s + 1)}{s(2s + 1)}, & G_3(s) &= \frac{1}{(s + 1)(0.04s^2 + 0.04s + 1)}, \\ G_d(s) &= \frac{-0.25(-s + 1)}{s(s + 1)(2s + 1)}, & G_4(s) &= \frac{1}{(s + 1)^4}, & G_5(s) &= \frac{1}{(0.2s + 1)^2} e^{-s}. \end{aligned} \quad (4.1)$$

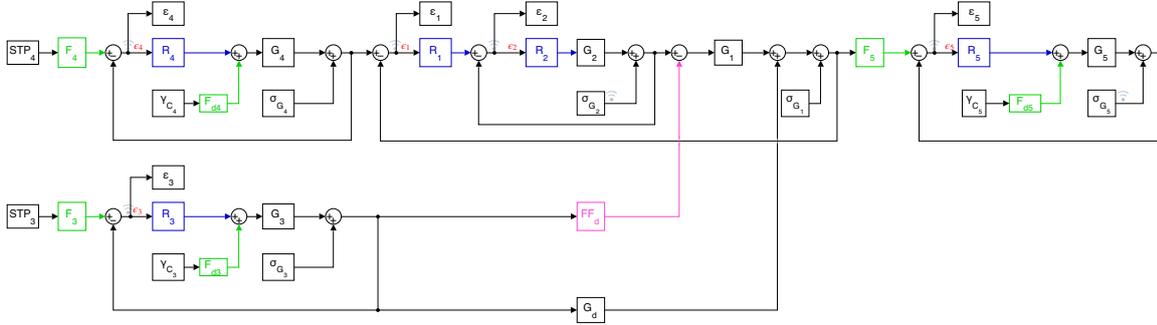


FIGURE 4.1: Simulation environment presenting multi-loop PID-based control layout (Falkowski and Domański, 2023)

Feedforward filters that are used to switch setpoint values are defined as:

$$F_3(s) = \frac{1}{4s^2 + 3.2s + 1}, \quad F_4(s) = \frac{1}{s^2 + 1.2s + 1}, \quad F_5(s) = \frac{1}{4s^2 + 2s + 1}. \quad (4.2)$$

Three filters are used to shape the fat-tailed disturbances:

$$F_{d3}(s) = \frac{1}{2s + 1}, \quad F_{d4}(s) = \frac{0.1}{s + 1}, \quad F_{d5}(s) = \frac{4}{2s + 1}. \quad (4.3)$$

Actually, the perfect or optimal tuning is not a goal of the considered research as the analysis aims at the propagation of the disturbances in the multi-loop structure. However, one has to be noticed, that the real industrial loops are seldom well-tuned (Ender, 1993), and therefore it is better to leave the loops just tuned without any significant efforts spent on tuning optimally. Controllers are tuned with the parameters sketched in Table 4.1.

TABLE 4.1: Tuning parameters of simulated controllers R_i – Matlab PID implementation in a pararel formulation

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
P	0.2	2	0.13	0.27	0.01
I	0.01	0.2	0.5	0.61	0.2
D	0.02	–	0.08	0.21	–
N	10	–	10	100	–

Due to the occurrence of disturbances in the proposed simulation system, disturbance decoupling is implemented using the industrial design (Domanski, 2020) with a block diagram and parameters sketched in Figure 4.2.

To simulate the opposite signal as close as possible to the disturbance in order to compensate it, two first-order inertial terms are used (with time constants T_1 and

4.1. Description of the simulation system

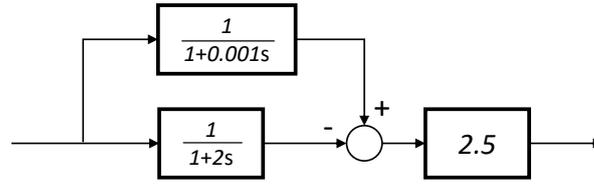


FIGURE 4.2: Layout of the industrial realization of the feedforward disturbance decoupling

T_2), which are subtracted from each other. The resulting signal is multiplied by the gain constant K . In order to select the appropriate parameters T_1 , T_2 , and K , the following method can be used (Trewiński, 2018):

1. Identify the disturbance signal and determine its frequency range and amplitude;
2. Choose a value for T_1 that is greater than the dominant frequency of the disturbance signal. This will ensure that the first inertial term has a sufficient time constant to filter out the disturbance;
3. Choose a value for T_2 that is less than the dominant frequency of the disturbance signal, but greater than the bandwidth of the system under control. This will ensure that the second inertial term responds quickly to changes in the system and helps compensate for the disturbance;
4. Adjust the gain constant K to achieve the desired level of compensation for the disturbance;
5. Test the system with the chosen parameters and adjust as necessary.

It is important to note that this method is not a "one-size-fits-all" solution and may need to be adjusted for specific systems and disturbances. Additionally, it is crucial to carefully consider the effects of adding the compensating signal to the system to ensure stability and avoid unintended consequences. Finally, the selected parameters are as follows: $T_1 = 0.001$, $T_2 = 2.0$, and $K = 2.5$.

The generated simulation data with added Gaussian noise are shown in Figure 4.3. The Gaussian noise parameters for each variable σ_i are presented in Table 4.2. The purpose of adding Gaussian noise is to simulate the inherent variability that is present in real-world data. This variability can arise due to several factors, such as measurement errors, environmental factors, or natural variations in the system.

Figure 4.4 presents the results of a simulation, where both Gauss and Cauchy disturbances are included. Parameters of each γ_{C_i} with sinusoidal signal of frequency equals 30 Hz are given in Table 4.2. The purpose of this simulation is to further investigate the impact of noise and disturbances on the causality results. By

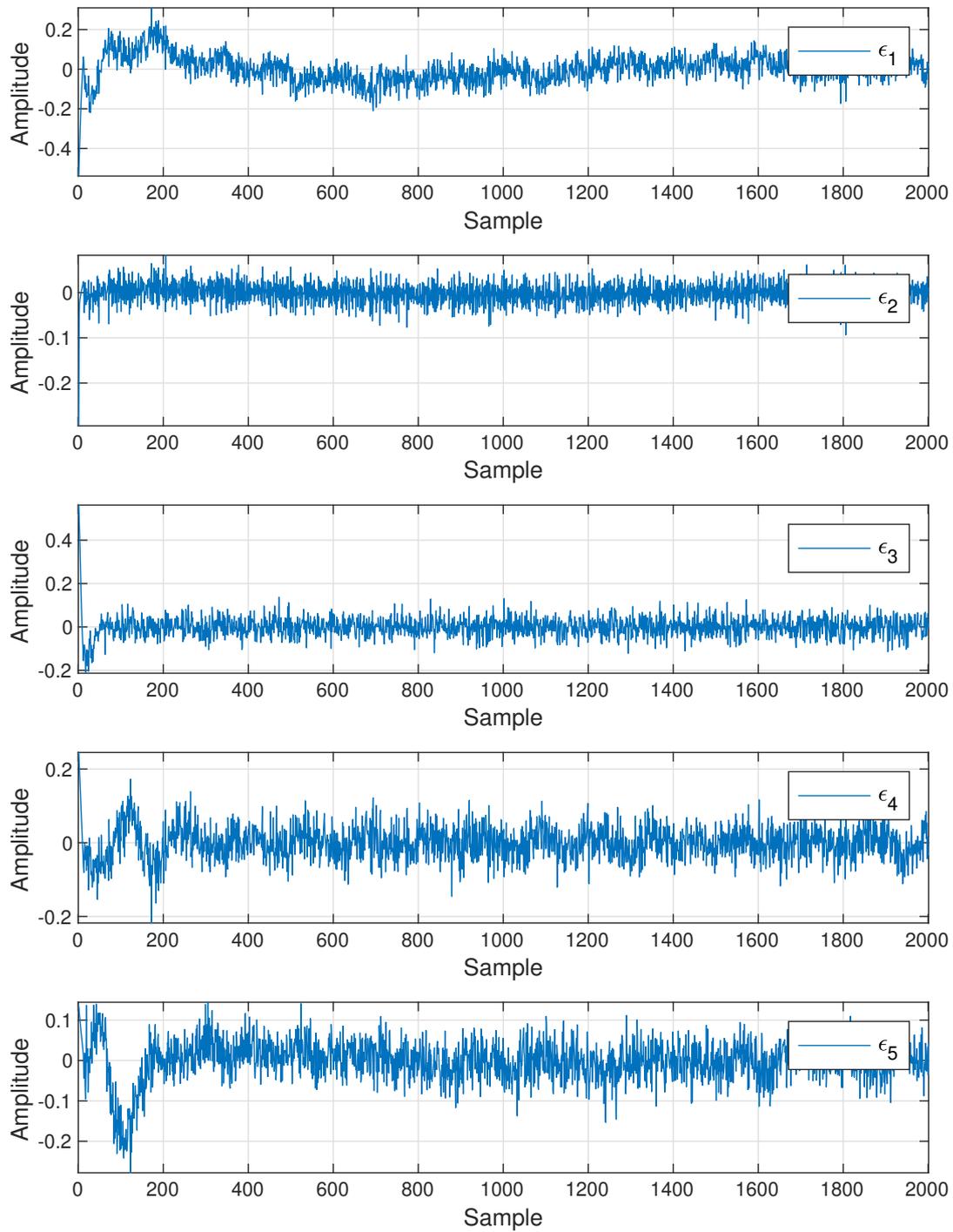


FIGURE 4.3: Control errors time series for the simulation data including Gaussian noise

4.1. Description of the simulation system

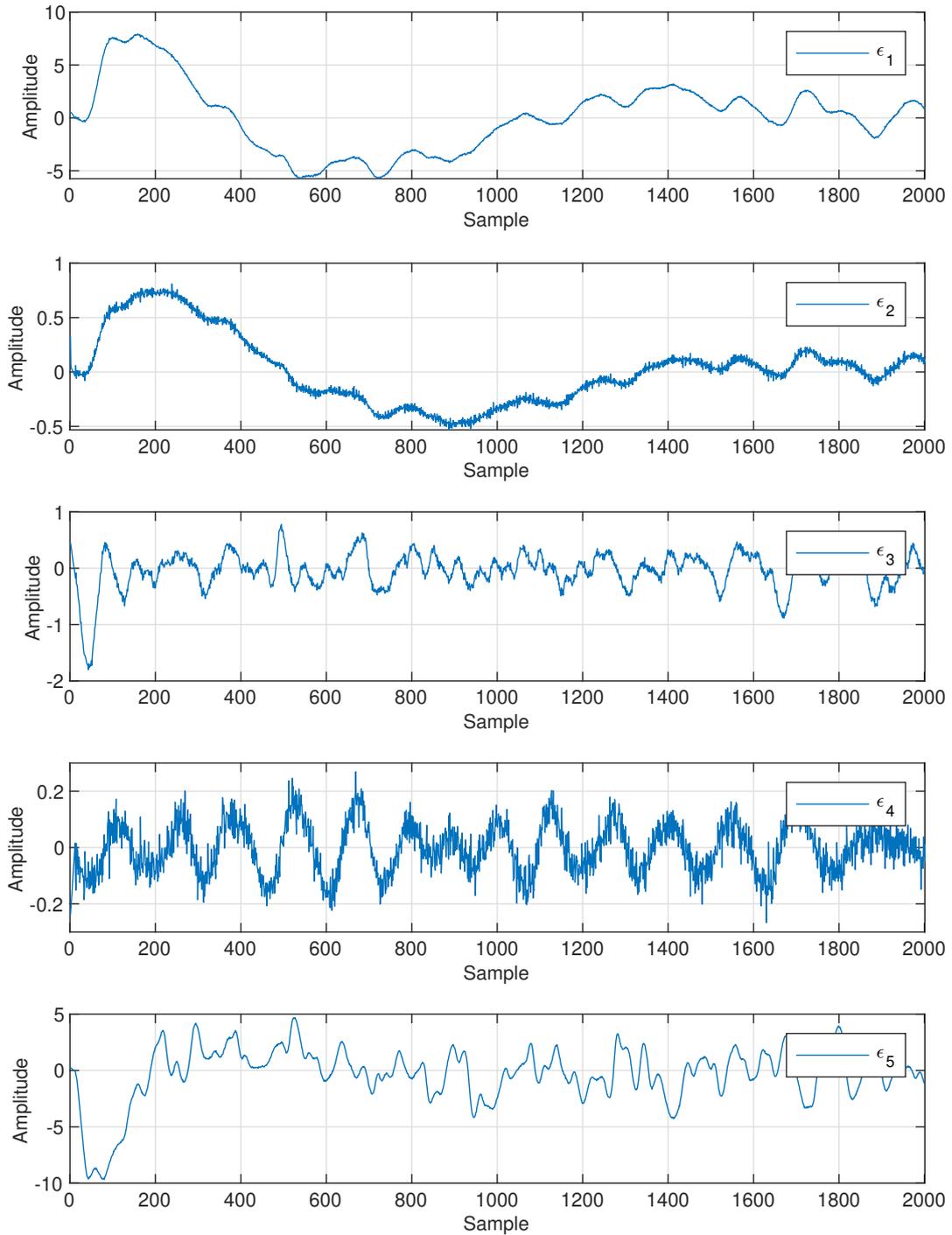


FIGURE 4.4: Control errors time series for the simulation data including Gaussian noise and Cauchy disturbance

TABLE 4.2: Gaussian noise and Cauchy disturbance signals parameters

	minimum	maximum	mean	median	standard deviation
σ_{G_1}	0.6444	1.4090	0.9989	1.0040	0.0975
σ_{G_2}	0.7030	1.3620	1.0020	1.0010	0.0991
σ_{G_3}	0.5751	1.3290	0.9983	0.9975	0.1018
σ_{G_4}	0.6910	1.3420	0.9994	1.0000	0.0998
σ_{G_5}	0.6939	1.3570	0.9988	0.9995	0.0985
γ_{C_3}	-0.6392	6.8330	3.0230	3.0240	1.2280
γ_{C_4}	-0.8589	6.7640	3.0310	2.9890	1.2190
γ_{C_5}	-1.0990	7.1130	2.9910	3.0430	1.2480

combining different types of noise and disturbances, a more complex and realistic simulation that better reflects the behavior of real-world systems is evaluated.

Root-cause relations between control errors are known, due to the known multi-loop structure. The true causality diagram for the simulation model is shown in Figure 4.5. The solid line denotes a direct relationship and the dashed line represents an indirect relationship.

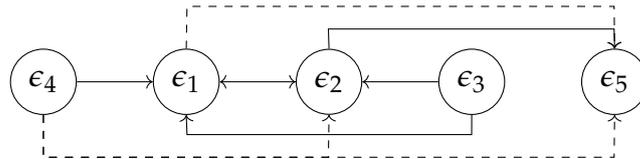


FIGURE 4.5: The actual causality diagram of the simulated benchmark

Histograms are a common tool used in data analysis to visualize the distribution of a dataset. In the context of control error histograms, they are used to observe the statistical properties of the data and to evaluate the effectiveness of the control system. The histograms presented in Figure 4.6, Figure 4.7, Figure 4.8, Figure 4.9, and Figure 4.10 show simulation data including Gaussian noise with PDF fitting.

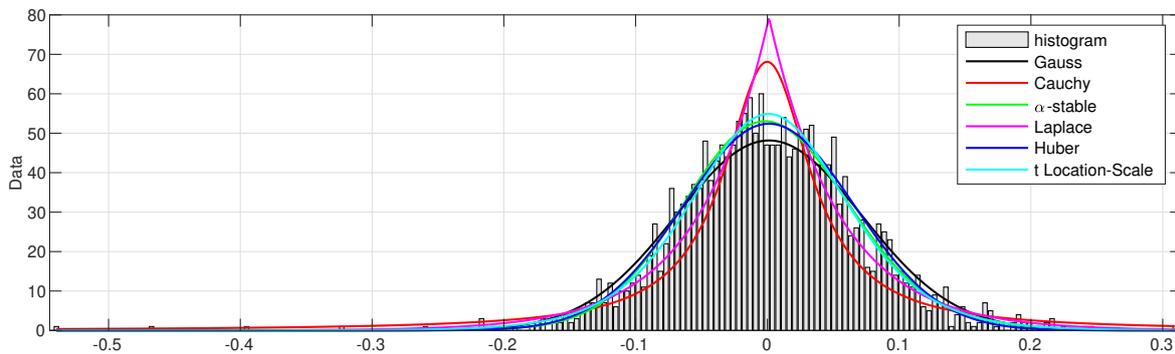


FIGURE 4.6: Control error ϵ_1 histogram for the simulation data including Gaussian noise with PDF fitting

4.1. Description of the simulation system

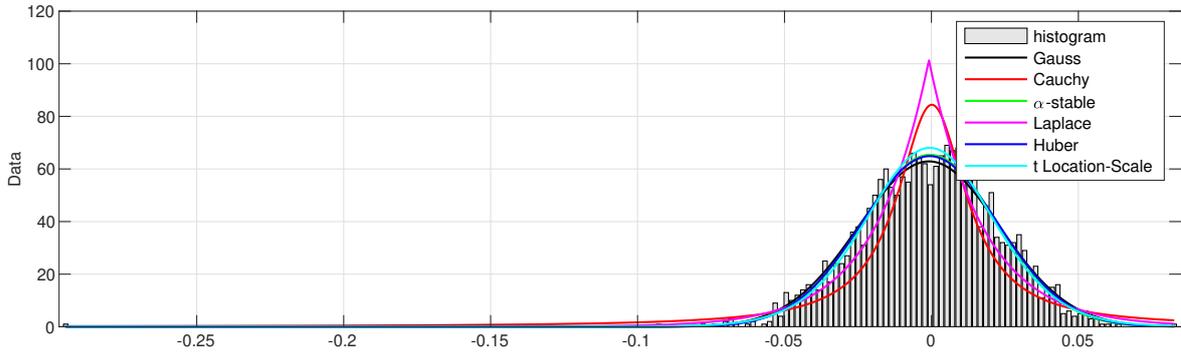


FIGURE 4.7: Control error ϵ_2 histogram for the simulation data including Gaussian noise with PDF fitting

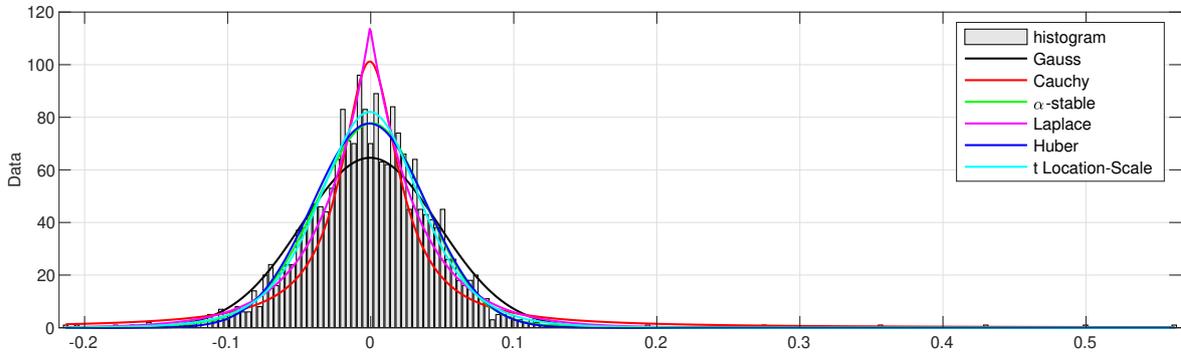


FIGURE 4.8: Control error ϵ_3 histogram for the simulation data including Gaussian noise with PDF fitting

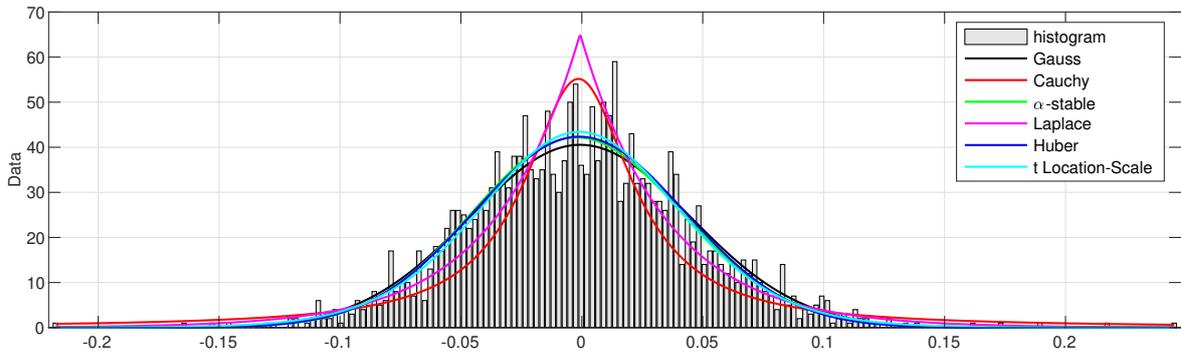


FIGURE 4.9: Control error ϵ_4 histogram for the simulation data including Gaussian noise with PDF fitting

For comparison purposes, apart from the Gaussian distribution, other distributions are also presented, i.e. Cauchy, α -stable, Laplace, Huber, and t Location-Scale. The presence of fat tails in the loop statistical properties suggests that the data has a higher likelihood of experiencing extreme events than what would be expected under a normal distribution. When attempting to fit a normal curve to data with fat

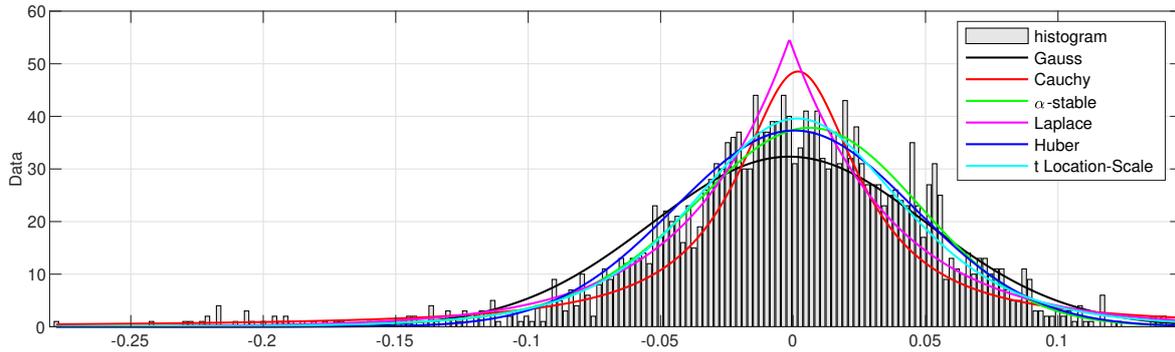


FIGURE 4.10: Control error ϵ_5 histogram for the simulation data including Gaussian noise with PDF fitting

tails, the curve underestimates the probability of extreme events and overestimates the probability of events in the central part of the distribution. This can lead to a misinterpretation of the data and an overestimation of the variance.

To avoid misfitting and variance overestimation, robust statistical scale estimators should be used. These estimators are designed to be less sensitive to outliers and extreme events, which can cause standard estimators such as the sample standard deviation to be biased. Examples of robust scale estimators include the median absolute deviation (MAD) and the interquartile range (IQR). Using robust statistical scale estimators provides a more accurate understanding of the data's statistical properties and may improve the effectiveness of the Transfer Entropy approach. It is important to carefully consider the choice of scale estimator based on the characteristics of the data and the specific application of the control system.

The histograms presented in Figure 4.11, Figure 4.12, Figure 4.13, Figure 4.14, and Figure 4.15 show simulation data that includes both Gaussian noise and Cauchy disturbance with PDF fitting. The histograms reveal that the data is not normally distributed and that there are fat tails in the control error data histograms.

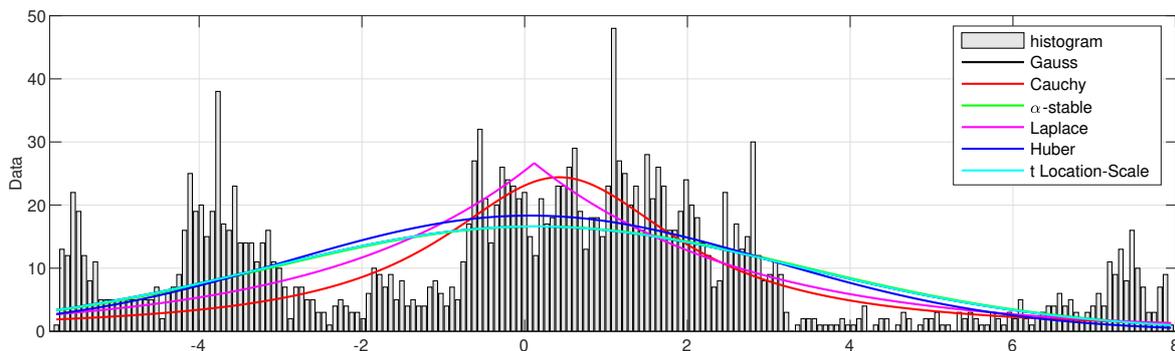


FIGURE 4.11: Control error ϵ_1 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting

4.1. Description of the simulation system

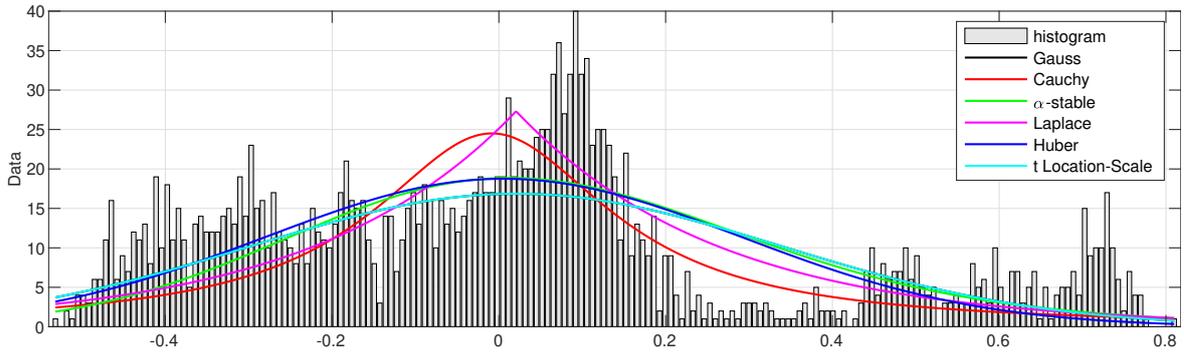


FIGURE 4.12: Control error ϵ_2 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting

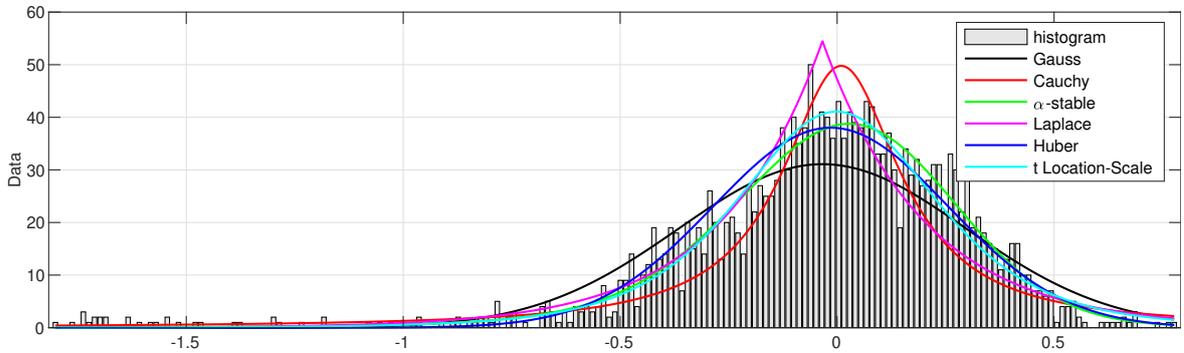


FIGURE 4.13: Control error ϵ_3 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting

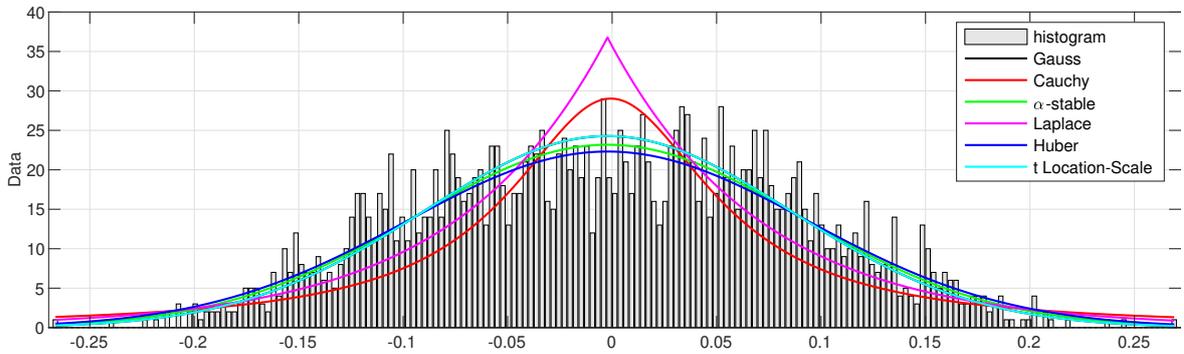


FIGURE 4.14: Control error ϵ_4 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting

Furthermore, it is shown that none of the presented distributions satisfactorily fits the data due to its irregular and scattered nature. This fact highlights the importance of taking into account alternative and robust statistical methods, once analyzing given datasets. In particular, heavy-tailed distributions like the Cauchy distribution may need to be considered, when using data-driven methods like Transfer

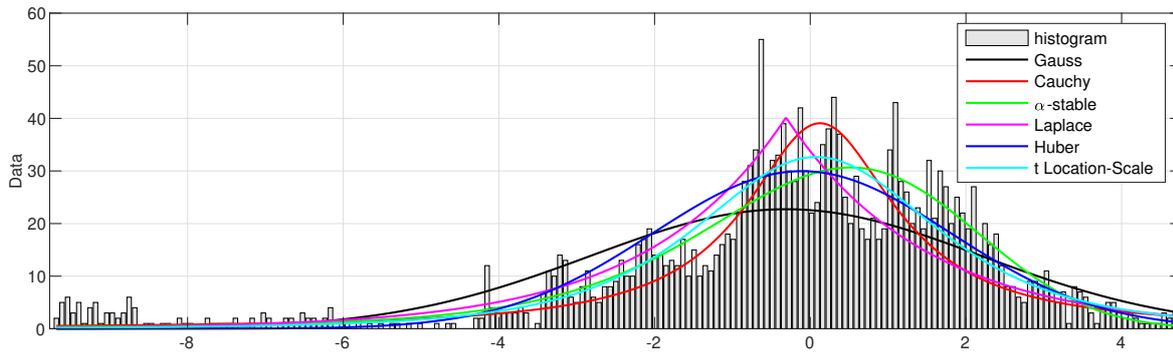


FIGURE 4.15: Control error ϵ_5 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting

Entropy.

Generally, the histograms of the simulation data reveal important information about the statistical properties of the control error datasets. They highlight the need for caution when assuming normality and emphasize the importance of using robust statistical methods to accurately fit the data.

4.2 Description of the real industrial system

Root cause analysis using the Transfer Entropy approach is employed to investigate the relationships in the ammonia production process at Grupa Azoty Zakłady Azotowe "Puławy" SA. The actual data collected from the ammonia plant is utilized for this purpose. The production of ammonia involves the autothermal reforming of methane (CH_4) using both pure oxygen and oxygen derived from the air (see Figure 4.16).

To prepare hydrogen for subsequent ammonia synthesis processing, the following sub-processes are involved:

1. Methane conversion (raw materials heating, including natural gas, process air, oxygen, and 3.2 MPa steam, in the preheaters, followed by autothermal methane reforming,
2. Carbon oxide conversion in shift reaction,
3. Removal of CO_2 through absorption in a hot potassium carbonate and activator solution in the Benfield unit,
4. Elimination of residual CO and CO_2 from the process gas through Copper-Ammonia Cleaning.

The initial stage of the process involves the conversion of methane into hydrogen, concurrently producing carbon monoxide (CO) and carbon dioxide (CO_2) as

4.2. Description of the real industrial system

secondary products, as outlined in the subsequent chemical reactions (Dziuba et al., 2020):

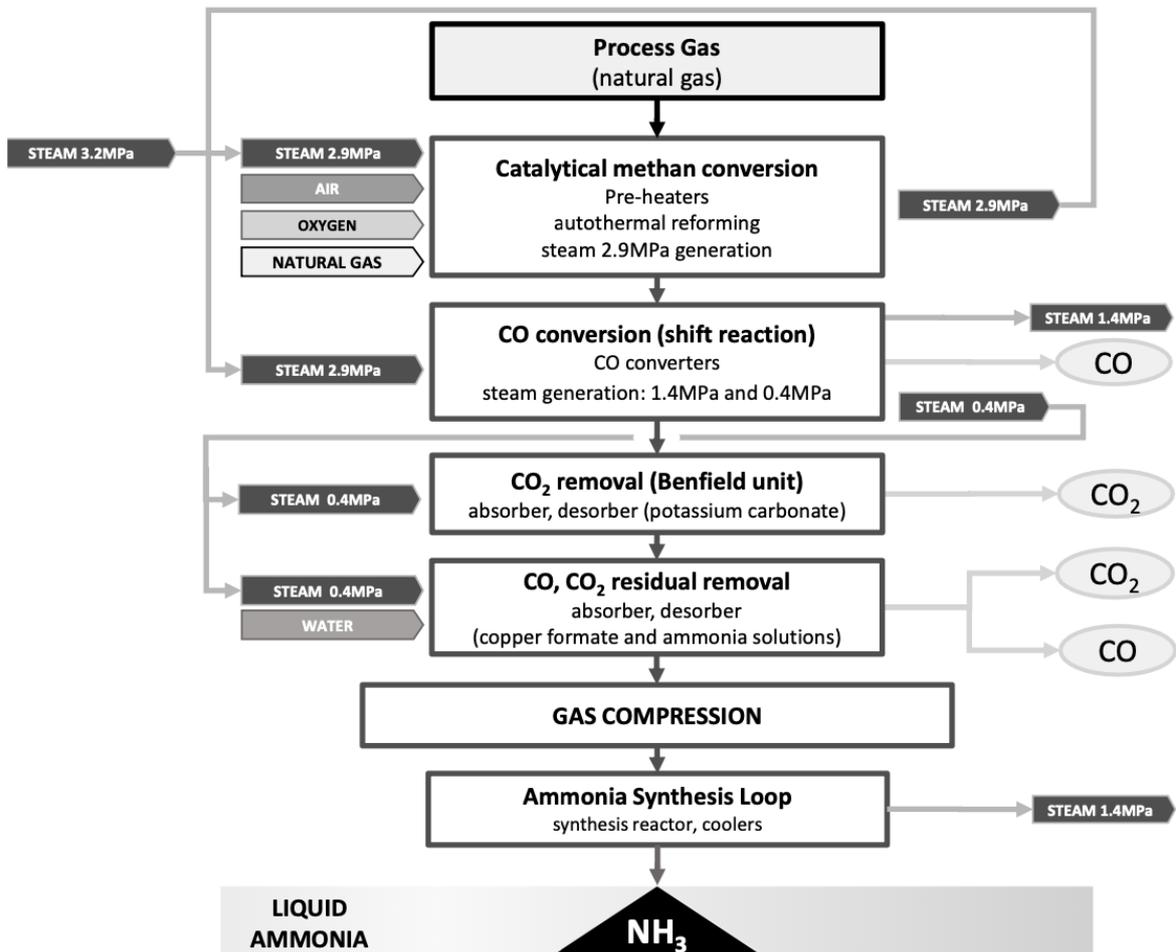
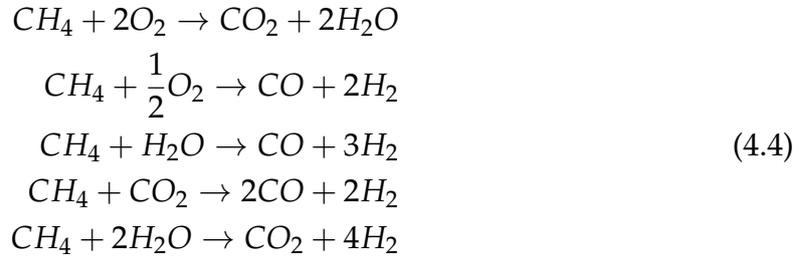


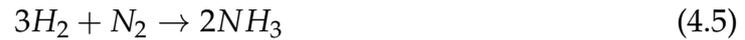
FIGURE 4.16: Ammonia production plant layout (Domański et al., 2023)

Following the initial conversion of methane into hydrogen, which generates undesired carbon oxide byproducts, the subsequent steps of the process involve further transformations. In the shift conversion section, carbon oxides are reacted with

process steam to produce additional hydrogen and carbon dioxide. To remove carbon dioxide from the syngas stream, absorption in a potassium carbonate solution (utilizing Benfield technology) is employed, and the separated CO_2 is redirected for urea production.

The final phase in the synthesis gas production process involves Copper-Ammonia Cleaning, which serves to eliminate any residual traces of carbon monoxide or carbon dioxide from the gas stream. To ultimately yield the desired end-product, ammonia, hydrogen is subjected to a catalytic reaction with nitrogen, which is sourced from process air. This reaction results in the formation of anhydrous liquid ammonia. The ammonia synthesis process occurs within a balanced system known as the synthesis loop, where circulating synthesis gas (syngas) is mixed with fresh input. Notably, the ammonia synthesis reaction transpires on a ferric catalyst, concurrently producing heat that is harnessed for steam generation and gas preheating.

The catalytic reaction between hydrogen and nitrogen can be described by the following chemical equation:



The entire process operates at a pressure of approximately 28-30 MPa. After the reaction, the produced ammonia is condensed through heat exchangers and chillers. The condensed ammonia is then collected in separators, where it undergoes decompression before being directed to subsequent installations for further processing.

For the purpose of conducting a causal analysis utilizing the Transfer Entropy method, a dataset derived from the ammonia production process, as elaborated upon earlier, is employed. This dataset is graciously provided by Grupa Azoty Zakłady Azotowe "Puławy" SA and encompasses control errors of flow (marked as ϵ_{F_i}), level (marked as ϵ_{L_i}), pressure (marked as ϵ_{P_i}), and temperature (marked as ϵ_{T_i}). Control errors are stemming from 22 intricately linked control loops within the overarching system (Domański et al., 2023). Furthermore, supplementary data pertaining to the environmental conditions surrounding the system is possessed. This dataset encompasses parameters such as ambient air temperature (T_α), external atmospheric pressure (P_e), air humidity levels (H), and air density (ρ_α) characteristics. This data will undergo comprehensive analysis to ascertain their influence on the causal relationships emerging from the interplay of control errors. This examination seeks to explore the potential impact of environmental factors on the causality analysis within the system.

The dataset is segregated into distinct subsets corresponding to 14 consecutive months. Unlike the simulation system, due to the intricate nature of this dataset, it is not feasible to straightforwardly depict the waveforms of individual control error signals in a manner that is easily interpretable and user-friendly. To provide an overview of the dataset's fundamental attributes, Figure 4.17 displays the temporal profiles of control errors in flow (ϵ_{F_7}), level (ϵ_{L_1}), pressure (ϵ_{P_1}), and temperature (ϵ_{T_2}). This data pertains specifically to the 8th month.

4.2. Description of the real industrial system

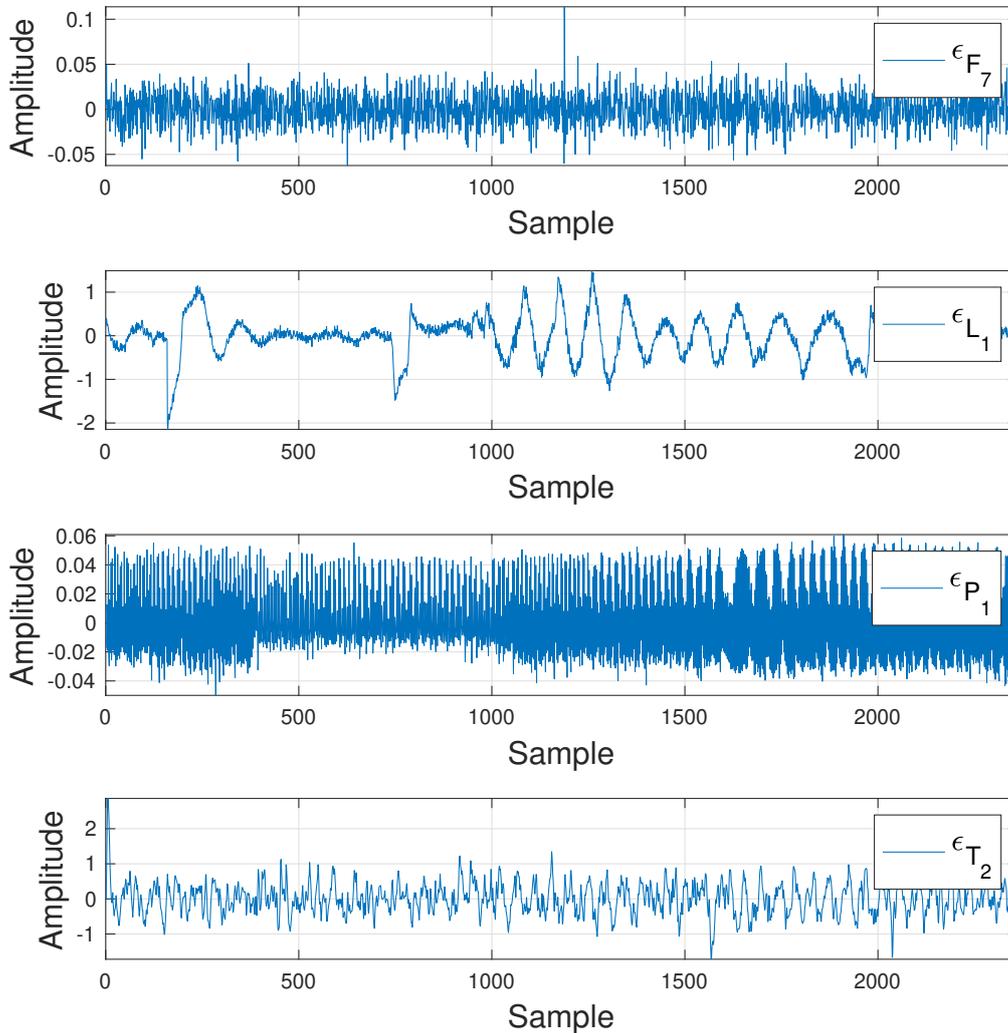


FIGURE 4.17: Selected control errors of flow, level, pressure, and temperature of the ammonia synthesis installation - dataset for the 8th month of operation

A thorough visual examination of the chosen control errors within the ammonia synthesis installation provides compelling evidence of its well-tuned nature. This assertion is substantiated by the absence of discernible oscillations in the characteristics of selected control errors.

Additionally, Figure 4.18 illustrates the ambient (outdoor) parameters corresponding to the same selected month. These parameters encompass essential environmental factors, such as ambient air temperature, air density, humidity level, and atmospheric pressure, which are pertinent to the broader context of the study.

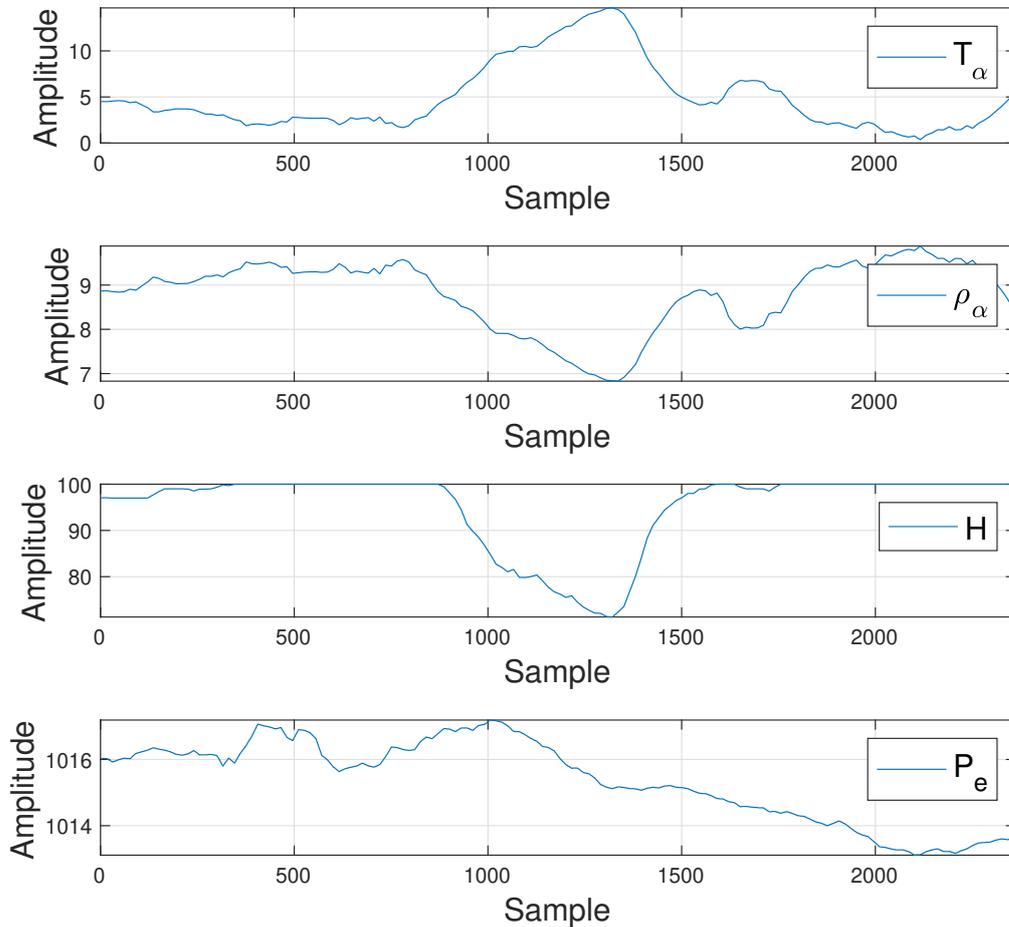


FIGURE 4.18: Selected weather parameters of ambient air temperature, air density, humidity level, and atmospheric pressure of the ammonia synthesis installation - dataset for the 8th month of operation

The characteristics of the weather data during the 8th month of the installation's operation align with anticipated patterns. Evident fluctuations in ambient temperature are observed, and the inverse relationship between humidity and rising ambient temperature validates the integrity and comprehensiveness of the collected dataset. It is noteworthy that the availability of 14 distinct datasets ensures coverage across all four seasons. It is essential to bear in mind that seasonal variations can lead to fluctuations in the contribution of weather data to the causality relationships associated with control errors. Moreover, the ambient conditions affect the process, as it uses air, whose density depends on weather conditions.

To gain a deeper comprehension of the data that will be utilized for causality analysis, there are generated probability distribution diagrams. Specifically, Figure

4.2. Description of the real industrial system

4.19 illustrates the probability distribution for flow control error (ϵ_{F_7}), Figure 4.20 depicts the distribution of level control error (ϵ_{L_1}), Figure 4.21 showcases the probability distribution of pressure control error (ϵ_{P_1}), and Figure 4.22 provides insight into the distribution of temperature control error (ϵ_{T_2}). Each of the figures presented illustrates the probability distributions for the respective variable over a span of 14 months. These distributions encapsulate the statistical characteristics and variations observed in the data across this multi-month timeframe, offering a comprehensive view of the data's stochastic behavior.

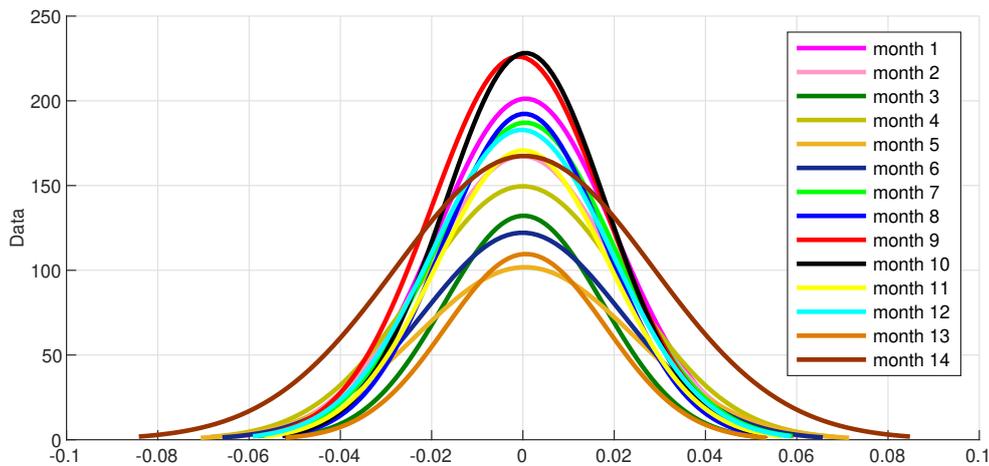


FIGURE 4.19: Probability distribution of the flow control error (ϵ_{F_7}) over a 14-month operational period of the ammonia synthesis installation

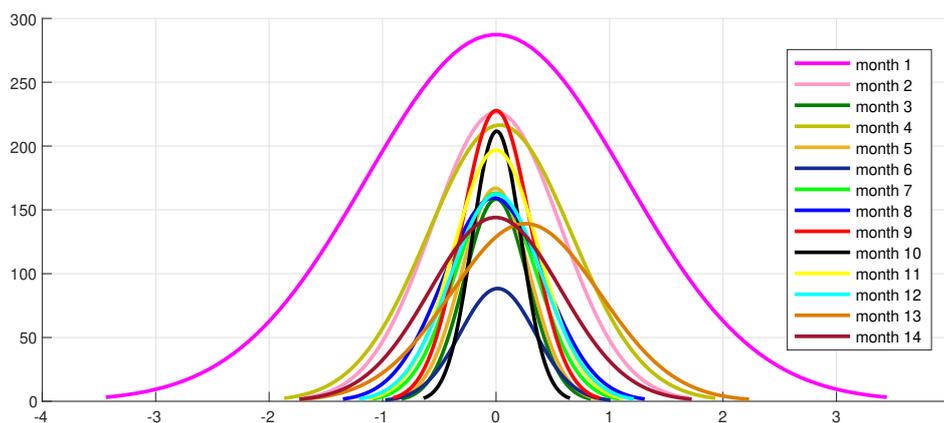


FIGURE 4.20: Probability distribution of the level control error (ϵ_{L_1}) over a 14-month operational period of the ammonia synthesis installation

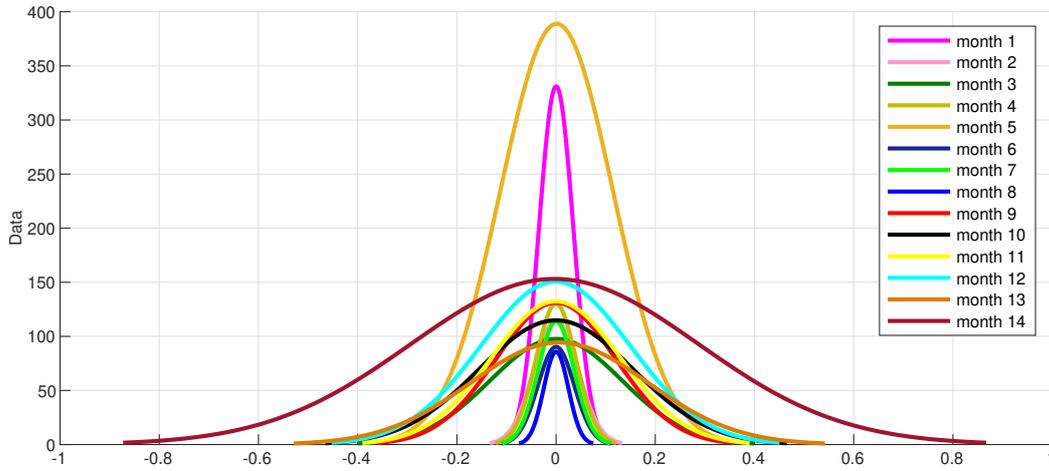


FIGURE 4.21: Probability distribution of the pressure control error (ϵ_{P_1}) over a 14-month operational period of the ammonia synthesis installation

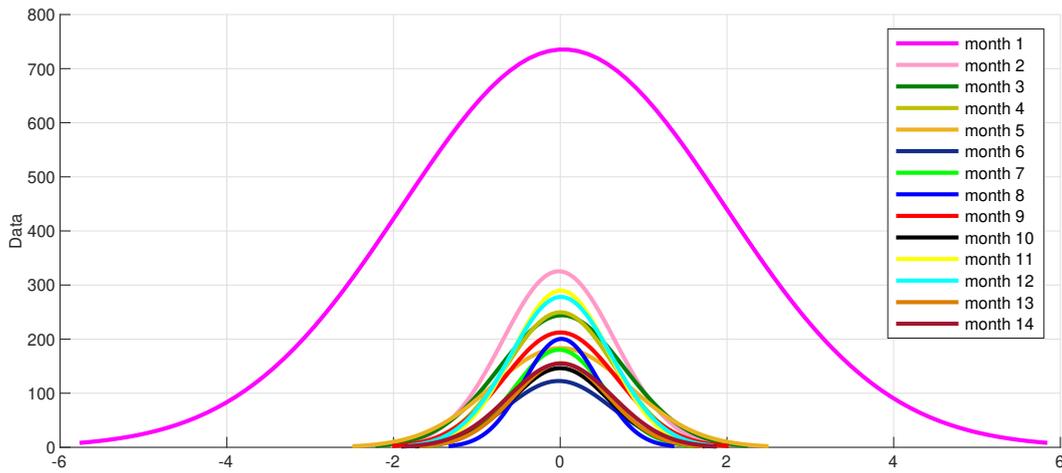


FIGURE 4.22: Probability distribution of the temperature control error (ϵ_{T_2}) over a 14-month operational period of the ammonia synthesis installation

In a majority of instances, the data distributions under examination exhibit characteristics akin to the normal distribution. This is particularly advantageous when employing the Transfer Entropy method, as its fundamental version relies on the assumption of normality. Nevertheless, in several scenarios where deviations from the normal distribution are observed, they do not pose significant impediments to our analytical pursuits.

Of paramount importance is the absence of outliers in the data representations. This absence can be attributed to the well-optimized performance of the system, thereby streamlining the subsequent data analysis procedures.

The conducted visual observations, represented through time-series charts of

4.2. Description of the real industrial system

specific control errors and probability distributions, while informative, do not offer a comprehensive overview of the whole dataset. Consequently, it is decided to determine Gaussian statistical metrics for the complete dataset (for all 14 months), categorized according to individual control errors. A comprehensive summary of these metrics, encompassing the mean, median, standard deviation, kurtosis, skewness, and median absolute deviation, are presented in Table 4.3, Table 4.4, Table 4.5, Table 4.6, Table 4.7, and Table 4.8, respectively. This approach aids in providing an understanding of the underlying characteristics of the dataset.

Analyzing the Gaussian statistical metrics for the control errors in the ammonia production process yields several noteworthy findings:

1. **Mean:** For ϵ_{F_1} to $\epsilon_{F_{12}}$, the mean values range from -0.17 to 0.26. It is notable that for most cases, control errors ϵ_{F_6} to $\epsilon_{F_{10}}$ have mean values equal to zero, suggesting that there is no significant deviation for these errors during the entire period. This may indicate the stable performance of control loops. Control errors ϵ_{F_8} and $\epsilon_{F_{11}}$ exhibit both positive and negative mean values, indicating fluctuations in performance over the months.

ϵ_{L_1} to ϵ_{L_3} show mean values close to zero. ϵ_{L_1} has a particularly high mean value in the 14th month (0.26), indicating a substantial deviation from the desired performance.

In the case of ϵ_{P_1} and ϵ_{T_1} to ϵ_{T_6} , most of these have mean values close to zero, suggesting that pressure and temperature deviations are relatively small and balanced over time.

2. **Median:** The median values for control factors range from -0.22 to 0.20. Similar to the mean values, control errors ϵ_{F_6} to $\epsilon_{F_{10}}$ manifest values equal to zero, suggesting that there are no significant deviations for these errors during the entire period. Control error ϵ_{F_2} has high-magnitude median values, indicating substantial variations and deviations over the months.

The median values for ϵ_{L_1} to ϵ_{L_3} are close to zero, indicating that the deviations are relatively balanced over time.

In the case of ϵ_{P_1} and ϵ_{T_1} to ϵ_{T_6} , most of these control errors have median values close to zero, suggesting that pressure and temperature deviations are typically small and balanced over time.

3. **Standard Deviation:** There is considerable variation in the standard deviation values across different control errors. Smaller standard deviations, exemplified by ϵ_{F_9} and $\epsilon_{F_{10}}$, indicate lower data variability, while larger values, seen in ϵ_{F_1} to ϵ_{F_3} , point to heightened variability within these errors. Standard deviation is responsible for the vertical extent of a probability distribution. It is primarily determined by the sample size unless the data is subject to normalization, which could potentially offer a more advantageous approach.

4. **Kurtosis:** Values greater than 3 indicate leptokurtic distributions, meaning they have heavier tails and a more peaked central region compared to a normal distribution (which has a kurtosis of 3). Control errors like $\epsilon_{F_1}, \epsilon_{F_4}, \epsilon_{F_5}, \epsilon_{F_7}, \epsilon_{F_8}, \epsilon_{F_{10}}, \epsilon_{F_{11}}, \epsilon_{F_{12}}, \epsilon_{L_2}, \epsilon_{L_3}, \epsilon_{P_1}, \epsilon_{T_1}, \epsilon_{T_2}, \epsilon_{T_4}$, and ϵ_{T_6} exhibit this characteristic, suggesting that the errors in these parameters have distributions with more pronounced peaks.

Control errors like $\epsilon_{F_3}, \epsilon_{F_6}, \epsilon_{F_9}, \epsilon_{L_1}, \epsilon_{L_3}$, and ϵ_{T_5} exhibit significant kurtosis values, signifying the presence of heavier tails and greater susceptibility to extreme errors.

Parameters with exceptionally high kurtosis values, such as $\epsilon_{F_4}, \epsilon_{F_6}, \epsilon_{L_3}$, and ϵ_{T_2} , suggest the presence of outliers or extreme values in the control error data for certain months. These outliers can contribute to increased variability and non-normality in the distributions.

Kurtosis values below 3, as seen in $\epsilon_{F_2}, \epsilon_{F_5}$, and ϵ_{F_7} , indicate platykurtic distributions, which have lighter tails and flatter central regions compared to a normal distribution. This suggests that the control errors for these parameters may exhibit less extreme values and relatively more stable behavior.

5. **Skewness:** Control errors, such as $\epsilon_{F_2}, \epsilon_{F_4}, \epsilon_{F_5}, \epsilon_{F_7}, \epsilon_{F_8}, \epsilon_{F_9}, \epsilon_{F_{10}}, \epsilon_{F_{11}}, \epsilon_{F_{12}}, \epsilon_{L_1}, \epsilon_{L_2}, \epsilon_{L_3}, \epsilon_{P_1}, \epsilon_{T_1}, \epsilon_{T_2}, \epsilon_{T_3}, \epsilon_{T_4}$, and ϵ_{T_5} , exhibit skewness values close to zero. This suggests that the distributions of control errors for these parameters are approximately symmetric.

Control errors with positive skewness, such as $\epsilon_{F_1}, \epsilon_{F_3}, \epsilon_{F_5}, \epsilon_{F_6}, \epsilon_{L_1}, \epsilon_{L_2}, \epsilon_{P_1}, \epsilon_{T_1}, \epsilon_{T_2}$, and ϵ_{T_4} , have distributions that are skewed to the right. This indicates that they may have a longer right tail, with potential outliers or extremely positive values.

Some control errors, like ϵ_{F_6} and ϵ_{T_6} , exhibit extreme positive skewness values. This suggests that their control error distributions have highly pronounced right tails, indicating the presence of significant outliers or extreme positive errors during certain months.

6. **Mean absolute deviation:** Control errors with higher MAD values, such as $\epsilon_{F_2}, \epsilon_{F_3}, \epsilon_{F_4}, \epsilon_{F_5}, \epsilon_{L_1}, \epsilon_{L_2}, \epsilon_{P_1}, \epsilon_{T_1}, \epsilon_{T_4}$, and ϵ_{T_5} , exhibit larger dispersion in their control errors. This implies that the errors for these parameters tend to deviate significantly from their respective means, indicating potential variability or instability.

Extremely high MAD values, such as those observed for ϵ_{F_2} , may suggest the presence of outliers or extreme control errors during those specific months.

4.2. Description of the real industrial system

TABLE 4.3: Gaussian statistical metric (mean) of the ammonia synthesis installation control errors

	Month													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ϵ_{F_1}	0.13	0.05	-0.17	-0.02	-0.01	0.07	-0.01	0.15	-0.04	0.08	0.11	0.00	-0.01	-0.03
ϵ_{F_2}	0.25	0.08	0.26	0.29	-0.11	-0.26	0.09	0.15	-0.30	0.07	-0.07	0.02	0.21	0.18
ϵ_{F_3}	0.09	0.04	-0.02	0.02	0.12	0.09	0.08	-0.20	-0.04	0.02	0.09	0.10	0.13	0.17
ϵ_{F_4}	0.00	-0.01	0.00	0.01	0.06	-0.04	0.00	-0.02	-0.02	0.00	0.02	0.00	0.00	0.00
ϵ_{F_5}	-0.01	-0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.02	-0.02
ϵ_{F_6}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ϵ_{F_7}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ϵ_{F_8}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.03	0.00	-0.07	-0.14	-0.22	-0.07	-0.20
ϵ_{F_9}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\epsilon_{F_{10}}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\epsilon_{F_{11}}$	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00	-0.01	0.00	0.02	0.00
$\epsilon_{F_{12}}$	0.02	-0.01	0.00	-0.03	0.04	0.02	0.02	0.01	-0.02	0.03	0.00	-0.02	0.02	0.02
ϵ_{L_1}	0.00	0.00	0.00	0.03	0.00	0.01	0.00	-0.02	0.00	0.01	0.00	0.00	0.26	0.00
ϵ_{L_2}	-0.01	0.00	-0.03	0.02	0.00	-0.12	0.00	0.00	-0.02	0.00	0.01	0.00	-0.03	0.00
ϵ_{L_3}	0.00	0.00	0.00	0.01	0.00	-0.01	0.00	0.01	0.00	0.01	0.00	0.00	-0.01	-0.02
ϵ_{P_1}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
ϵ_{T_1}	0.06	-0.02	-0.13	-0.02	0.10	-0.01	0.02	0.00	0.03	0.00	0.01	-0.02	-0.04	0.04
ϵ_{T_2}	0.04	-0.02	0.02	0.00	0.00	-0.02	-0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.00
ϵ_{T_3}	-0.02	0.00	-0.01	0.00	-0.03	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.01
ϵ_{T_4}	0.03	0.00	0.02	-0.01	-0.07	-0.05	0.01	0.00	-0.03	-0.01	0.01	-0.04	0.00	0.02
ϵ_{T_5}	-0.01	0.01	-0.01	-0.01	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ϵ_{T_6}	0.01	0.02	0.02	0.00	0.00	0.02	0.00	0.02	0.02	0.00	0.00	0.01	0.01	0.00

TABLE 4.4: Gaussian statistical metric (median) of the ammonia synthesis installation control errors

	Month													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ϵ_{F_1}	-0.03	0.02	-0.22	0.04	0.04	0.04	-0.02	0.11	-0.04	-0.02	0.13	0.20	-0.01	-0.02
ϵ_{F_2}	0.31	-0.36	0.16	0.33	0.11	-0.47	-0.25	0.20	-0.29	-1.63	-0.22	-0.11	-0.18	0.20
ϵ_{F_3}	-0.05	-0.08	0.00	-0.14	0.08	-0.08	0.05	-0.26	0.09	0.01	-0.05	0.06	-0.07	0.17
ϵ_{F_4}	0.00	-0.03	0.01	0.19	0.05	-0.08	0.01	-0.01	-0.03	0.00	0.03	0.01	-0.02	0.01
ϵ_{F_5}	0.01	-0.02	-0.01	0.00	-0.01	0.01	0.01	0.00	0.01	0.01	0.03	0.00	0.02	-0.03
ϵ_{F_6}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
ϵ_{F_7}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ϵ_{F_8}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.03	0.00	-0.08	-0.15	-0.22	-0.07	-0.21
ϵ_{F_9}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\epsilon_{F_{10}}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\epsilon_{F_{11}}$	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	-0.01	0.00	-0.01	-0.01	0.01	0.00
$\epsilon_{F_{12}}$	0.02	-0.02	0.00	-0.03	0.04	0.00	0.02	0.00	-0.03	0.02	-0.01	-0.02	0.00	0.01
ϵ_{L_1}	0.01	-0.01	0.00	0.00	0.00	0.03	0.01	0.01	0.00	0.00	-0.01	0.01	0.02	0.00
ϵ_{L_2}	-0.03	0.01	-0.01	0.03	0.01	-0.06	0.00	-0.01	-0.01	0.01	0.01	0.00	-0.01	-0.03
ϵ_{L_3}	0.00	0.00	0.04	0.02	0.02	-0.03	-0.01	0.00	-0.01	-0.01	0.00	-0.01	-0.04	0.00
ϵ_{P_1}	-0.01	-0.01	-0.04	-0.01	-0.01	-0.01	-0.01	-0.01	0.02	0.01	0.02	0.03	0.05	-0.08
ϵ_{T_1}	0.01	-0.11	-0.14	-0.02	0.12	0.12	0.04	0.02	0.01	-0.04	-0.02	0.00	-0.13	0.02
ϵ_{T_2}	0.00	-0.04	-0.01	-0.02	0.03	-0.03	-0.02	0.00	-0.03	-0.01	-0.01	0.00	0.00	0.01
ϵ_{T_3}	0.00	-0.01	0.00	0.01	-0.04	0.00	-0.01	0.01	0.00	0.00	-0.01	-0.01	-0.02	-0.01
ϵ_{T_4}	-0.08	-0.04	0.02	0.05	0.03	-0.11	0.03	-0.03	0.01	-0.01	0.09	0.02	0.20	0.03
ϵ_{T_5}	0.00	0.00	0.00	0.00	0.04	-0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
ϵ_{T_6}	0.04	0.05	0.05	0.01	0.01	0.04	0.08	0.07	0.08	0.08	0.02	0.04	0.05	-0.01

4.2. Description of the real industrial system

TABLE 4.5: Gaussian statistical metric (standard deviation) of the ammonia synthesis installation control errors

	Month													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ϵ_{F_1}	5.20	4.89	3.78	5.25	6.45	6.21	5.23	5.08	5.16	5.06	4.60	6.16	4.20	3.40
ϵ_{F_2}	15.08	17.04	16.10	16.37	19.77	25.74	21.81	21.09	19.74	21.69	20.89	16.09	15.63	16.17
ϵ_{F_3}	7.07	6.24	7.63	7.35	8.23	10.44	7.65	10.69	9.68	8.42	7.46	6.49	6.19	6.01
ϵ_{F_4}	1.23	1.77	1.02	1.41	0.79	1.21	0.95	0.95	1.37	1.23	1.40	2.06	0.97	1.53
ϵ_{F_5}	0.87	0.47	0.44	0.51	0.43	0.39	0.36	0.39	0.54	0.57	0.68	1.08	0.99	0.70
ϵ_{F_6}	0.04	0.06	0.05	0.06	0.05	0.06	0.04	0.04	0.07	0.06	0.06	0.05	0.05	0.06
ϵ_{F_7}	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03
ϵ_{F_8}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.04	0.01	0.03	0.03
ϵ_{F_9}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\epsilon_{F_{10}}$	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.03
$\epsilon_{F_{11}}$	0.13	0.17	0.12	0.12	0.11	0.14	0.17	0.21	0.23	0.24	0.23	0.27	0.27	0.20
$\epsilon_{F_{12}}$	0.48	0.40	1.22	0.39	0.49	0.37	0.36	0.41	0.44	0.39	0.42	0.39	0.40	0.39
ϵ_{L_1}	1.15	0.57	0.28	0.63	0.30	0.33	0.36	0.44	0.30	0.21	0.37	0.40	0.66	0.58
ϵ_{L_2}	0.76	0.47	0.31	0.49	0.56	0.62	0.34	0.35	0.55	0.33	0.40	0.65	0.19	0.67
ϵ_{L_3}	0.76	0.43	0.24	0.20	0.29	0.39	0.28	0.25	0.44	0.22	0.23	0.17	0.31	0.61
ϵ_{P_1}	0.03	0.04	0.13	0.04	0.11	0.04	0.04	0.02	0.12	0.16	0.13	0.15	0.18	0.29
ϵ_{T_1}	2.03	1.91	2.30	1.56	2.07	2.17	1.59	0.69	1.53	0.90	0.81	1.19	2.01	1.20
ϵ_{T_2}	1.93	0.67	0.74	0.64	0.83	0.63	0.54	0.45	0.67	0.58	0.58	0.61	0.57	0.63
ϵ_{T_3}	0.44	0.45	0.52	0.49	0.58	0.42	0.39	0.39	0.45	0.43	0.44	0.44	0.43	0.66
ϵ_{T_4}	3.74	2.80	4.14	2.54	3.38	3.16	2.64	1.39	2.43	1.78	1.73	2.20	3.15	2.24
ϵ_{T_5}	0.13	0.30	0.15	0.34	0.29	0.21	0.11	0.11	0.13	0.10	0.27	0.12	0.11	0.73
ϵ_{T_6}	0.34	0.42	0.36	0.30	0.27	0.31	0.34	0.48	0.66	0.50	0.40	0.43	0.41	0.31

TABLE 4.6: Gaussian statistical metric (kurtosis) of the ammonia synthesis installation control errors

	Month													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ϵ_{F_1}	97.10	8.47	3.86	19.03	10.80	18.25	5.50	3.12	6.11	3.19	5.29	11.78	3.70	5.37
ϵ_{F_2}	58.21	19.94	7.03	21.05	12.22	103.55	5.86	46.27	5.40	6.82	9.31	8.57	5.14	15.76
ϵ_{F_3}	15.70	5.50	5.75	5.98	4.69	151.91	5.27	3.82	5.82	3.73	4.70	4.18	5.11	4.47
ϵ_{F_4}	481.10	3.84	5.18	3.14	3.41	3.01	3.21	5.50	3.31	3.00	3.23	3.23	3.94	3.78
ϵ_{F_5}	190.21	2.67	2.86	2.46	2.66	2.86	2.85	3.27	2.58	2.48	2.44	2.11	2.06	2.51
ϵ_{F_6}	7.05	4.59	3.87	3.15	4.29	3.08	3.60	3.18	266.86	3.09	88.28	3.20	2.79	3.30
ϵ_{F_7}	3.12	2.91	3.15	2.98	3.20	3.17	3.24	3.59	3.26	3.43	3.09	3.10	2.96	3.28
ϵ_{F_8}	25.29	3.32	3.60	20.95	14.12	3.31	2.65	2.86	3.23	2.51	4.70	13.18	1.71	2.60
ϵ_{F_9}	2.83	2.65	3.03	3.63	2.66	3.01	2.81	242.40	2.65	3.12	3.11	15.03	3.10	3.01
$\epsilon_{F_{10}}$	6.23	3.77	2.58	4.05	3.44	2.42	2.56	2.62	2.31	2.52	4.38	2.55	2.41	3.14
$\epsilon_{F_{11}}$	3.22	3.29	3.13	3.09	3.43	3.46	3.74	3.19	3.19	3.11	3.10	3.00	3.09	3.07
$\epsilon_{F_{12}}$	66.12	5.34	727.31	5.08	14.27	3.35	3.41	3.34	3.95	4.10	4.76	3.25	3.26	3.85
ϵ_{L_1}	17.98	6.06	7.65	12.52	15.45	2.79	3.89	4.87	6.38	3.40	5.38	3.24	5.13	4.49
ϵ_{L_2}	7.65	4.48	2.73	11.50	10.60	4.55	2.74	4.36	7.41	2.56	3.59	4.87	33.54	2.59
ϵ_{L_3}	7.97	26.81	4.77	3.38	3.47	10.57	5.19	33.60	34.87	17.17	1.95	19.41	69.75	441.16
ϵ_{P_1}	5.83	2.26	3.19	3.27	273.96	4.43	3.91	2.31	3.82	2.36	3.33	5.96	6.01	3.79
ϵ_{T_1}	35.83	9.63	8.72	15.04	4.89	6.15	6.63	5.51	18.88	4.01	3.79	5.53	7.77	9.34
ϵ_{T_2}	260.85	8.16	18.60	6.28	20.84	3.20	3.20	5.36	3.60	2.88	8.47	7.49	4.07	3.19
ϵ_{T_3}	18.55	3.58	13.14	3.60	8.66	3.59	3.51	6.08	4.13	3.82	4.23	5.34	4.29	4.51
ϵ_{T_4}	12.97	7.66	13.68	7.86	3.70	4.96	5.55	3.55	11.48	3.49	3.44	5.50	5.31	6.08
ϵ_{T_5}	3.17	5.91	5.72	3.34	4.16	2.68	3.25	3.27	3.64	3.19	4.94	10.73	4.83	2.12
ϵ_{T_6}	3.70	14.67	3.47	26.63	5.54	3.10	3.65	11.32	9.67	4.08	3.67	3.41	7.83	3.64

4.2. Description of the real industrial system

TABLE 4.7: Gaussian statistical metric (skewness) of the ammonia synthesis installations control errors

	Month													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ϵ_{F_1}	4.66	0.08	-0.09	-0.72	0.07	0.01	-0.14	0.04	0.31	0.07	0.22	-1.03	-0.09	-0.10
ϵ_{F_2}	1.81	0.75	0.11	-0.46	-0.73	4.17	0.61	-1.48	0.15	1.17	0.18	0.08	0.03	-0.20
ϵ_{F_3}	-0.01	0.08	-0.05	-0.09	-0.27	4.39	0.07	0.00	-0.06	0.00	0.12	0.05	0.18	0.13
ϵ_{F_4}	12.93	-0.01	0.11	-0.45	0.02	0.10	-0.10	-0.37	-0.05	-0.05	-0.11	-0.09	0.00	-0.10
ϵ_{F_5}	6.17	0.10	0.08	-0.02	0.16	0.06	0.07	0.11	0.04	0.00	-0.03	0.04	0.01	0.02
ϵ_{F_6}	-0.17	0.00	-0.28	0.06	0.25	-0.16	-0.10	0.31	-8.79	-0.27	-3.97	-0.22	-0.26	-0.08
ϵ_{F_7}	-0.01	-0.08	-0.17	-0.21	-0.22	-0.10	-0.10	0.01	0.06	-0.14	-0.04	-0.06	-0.09	-0.09
ϵ_{F_8}	0.40	-0.14	-0.37	0.27	0.16	-0.01	0.03	-0.12	-0.05	0.89	1.55	0.99	0.17	0.39
ϵ_{F_9}	-0.40	-0.38	-0.32	-0.33	-0.31	-0.28	-0.11	-8.20	-0.12	-0.28	-0.24	0.47	-0.12	0.09
$\epsilon_{F_{10}}$	-0.42	0.03	-0.12	-0.01	0.04	-0.09	-0.13	-0.27	-0.15	-0.17	0.09	-0.18	-0.16	-0.01
$\epsilon_{F_{11}}$	-0.05	0.17	0.05	0.06	0.12	0.09	0.26	0.18	0.18	0.12	0.11	0.09	0.11	0.14
$\epsilon_{F_{12}}$	-3.59	0.42	-24.31	-0.10	1.13	0.12	0.04	0.16	0.22	0.27	0.50	0.03	0.17	0.23
ϵ_{L_1}	0.15	0.32	0.78	0.88	-1.44	0.07	-0.19	-0.46	0.05	0.25	0.44	-0.11	0.00	0.17
ϵ_{L_2}	0.55	-0.29	-0.04	-0.72	-0.81	-0.83	-0.24	-0.42	-1.01	-0.13	0.16	0.00	-2.84	0.09
ϵ_{L_3}	-0.36	1.57	-0.13	0.21	-0.18	-0.65	0.34	3.41	1.85	1.58	-0.06	1.73	5.93	-17.80
ϵ_{P_1}	1.03	0.42	0.86	0.70	12.72	1.55	1.40	0.58	-1.01	-0.38	-0.89	-1.69	-1.72	0.56
ϵ_{T_1}	3.33	0.24	-0.34	0.58	0.17	0.01	-0.05	-0.27	-0.76	0.29	0.12	-0.19	0.63	0.60
ϵ_{T_2}	12.50	-0.27	2.11	0.29	2.11	0.12	0.13	0.43	0.22	0.13	0.70	0.64	0.21	0.11
ϵ_{T_3}	-1.70	0.09	-0.82	-0.18	-0.79	-0.05	0.03	-0.38	-0.16	0.00	-0.09	-0.25	-0.14	-0.05
ϵ_{T_4}	0.72	0.20	-0.01	0.20	-0.20	0.08	0.10	0.08	-0.48	0.06	-0.24	-0.51	-0.09	0.48
ϵ_{T_5}	-0.13	0.29	-0.24	-0.08	-0.88	0.05	0.18	0.11	-0.02	-0.28	-0.06	-1.24	-0.39	-0.05
ϵ_{T_6}	-0.48	1.27	-0.51	-2.82	0.20	-0.37	-0.60	0.71	-0.38	-0.80	-0.75	-0.50	-1.01	-0.03

TABLE 4.8: Gaussian statistical metric (mean absolute deviation) of the ammonia synthesis installation control errors

	Month													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ϵ_{F_1}	3.23	3.52	2.96	3.85	4.63	4.52	3.91	4.04	4.06	4.03	3.50	4.21	3.29	2.56
ϵ_{F_2}	10.05	11.03	11.65	11.82	14.12	16.78	15.82	14.88	15.06	15.65	14.83	11.97	11.66	10.91
ϵ_{F_3}	5.06	4.55	5.61	5.42	6.10	6.77	5.56	8.19	7.41	6.48	5.52	4.94	4.67	4.54
ϵ_{F_4}	0.71	1.31	0.76	1.12	0.61	0.97	0.75	0.67	1.08	0.99	1.12	1.58	0.74	1.13
ϵ_{F_5}	0.60	0.39	0.36	0.42	0.35	0.31	0.29	0.30	0.44	0.47	0.57	0.91	0.84	0.57
ϵ_{F_6}	0.03	0.04	0.04	0.05	0.04	0.04	0.03	0.03	0.05	0.05	0.04	0.04	0.04	0.04
ϵ_{F_7}	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.01	0.02	0.01	0.02
ϵ_{F_8}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.03	0.01	0.03	0.02
ϵ_{F_9}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\epsilon_{F_{10}}$	0.01	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
$\epsilon_{F_{11}}$	0.10	0.13	0.09	0.10	0.08	0.11	0.13	0.17	0.18	0.19	0.18	0.22	0.22	0.16
$\epsilon_{F_{12}}$	0.32	0.31	0.41	0.30	0.33	0.29	0.29	0.33	0.34	0.31	0.32	0.30	0.32	0.30
ϵ_{L_1}	0.50	0.40	0.18	0.33	0.16	0.26	0.28	0.32	0.21	0.17	0.26	0.31	0.50	0.41
ϵ_{L_2}	0.48	0.35	0.25	0.32	0.35	0.47	0.28	0.29	0.40	0.28	0.32	0.49	0.09	0.54
ϵ_{P_1}	0.03	0.04	0.11	0.03	0.05	0.03	0.03	0.02	0.09	0.13	0.10	0.10	0.12	0.22
ϵ_{T_1}	1.14	1.25	1.38	1.04	1.48	1.48	1.11	0.51	0.84	0.69	0.63	0.86	1.36	0.88
ϵ_{T_2}	0.58	0.50	0.50	0.49	0.58	0.50	0.43	0.35	0.53	0.46	0.43	0.46	0.45	0.50
ϵ_{T_3}	0.31	0.36	0.37	0.39	0.43	0.33	0.30	0.29	0.35	0.33	0.34	0.34	0.33	0.51
ϵ_{T_4}	2.42	1.90	2.45	1.83	2.59	2.30	1.92	1.09	1.58	1.40	1.35	1.63	2.28	1.73
ϵ_{T_5}	0.10	0.20	0.11	0.26	0.22	0.17	0.08	0.09	0.10	0.07	0.19	0.09	0.08	0.61
ϵ_{L_3}	0.45	0.21	0.19	0.17	0.24	0.28	0.22	0.13	0.22	0.15	0.19	0.11	0.13	0.17
ϵ_{T_6}	0.27	0.31	0.28	0.20	0.20	0.25	0.27	0.36	0.46	0.38	0.31	0.34	0.30	0.24

4.2. Description of the real industrial system

To summarize, the control errors demonstrate varying levels of data variability, tail heaviness or lightness, and skewness in either direction. These statistical metrics provide valuable insights into the underlying data characteristics, serving as a foundation for further causality analysis aimed at enhancing the Transfer Entropy approach effectiveness, which results in better reflection of relationships between control loops.

Considering a familiarity with the ammonia synthesis installation, Figure 4.23 displays the actual causality graph stemming from the interconnections among the control loops. This graph will serve as the foundational reference for evaluating the efficacy of the Transfer Entropy approach in scrutinizing causal relationships within the system. The solid line denotes a direct relationship and the dashed line represents an indirect relationship. It is imperative to note that Figure 4.23 exclusively showcases the depiction of control errors represented as nodes. Currently, it is not feasible to visually represent the variables associated with ambient (outdoor) parameters, specifically ambient temperature, air density, humidity level, and atmospheric pressure. This limitation arises from the absence of a clear understanding regarding the interrelationship between these ambient parameters and control errors.

The verification of the Transfer Entropy method's efficacy is expounded upon in Chapter 10. This validation process is rooted in a sequence of meticulously conducted investigations, the detailed chronological account and outcomes of which are comprehensively delineated in the ensuing chapters.

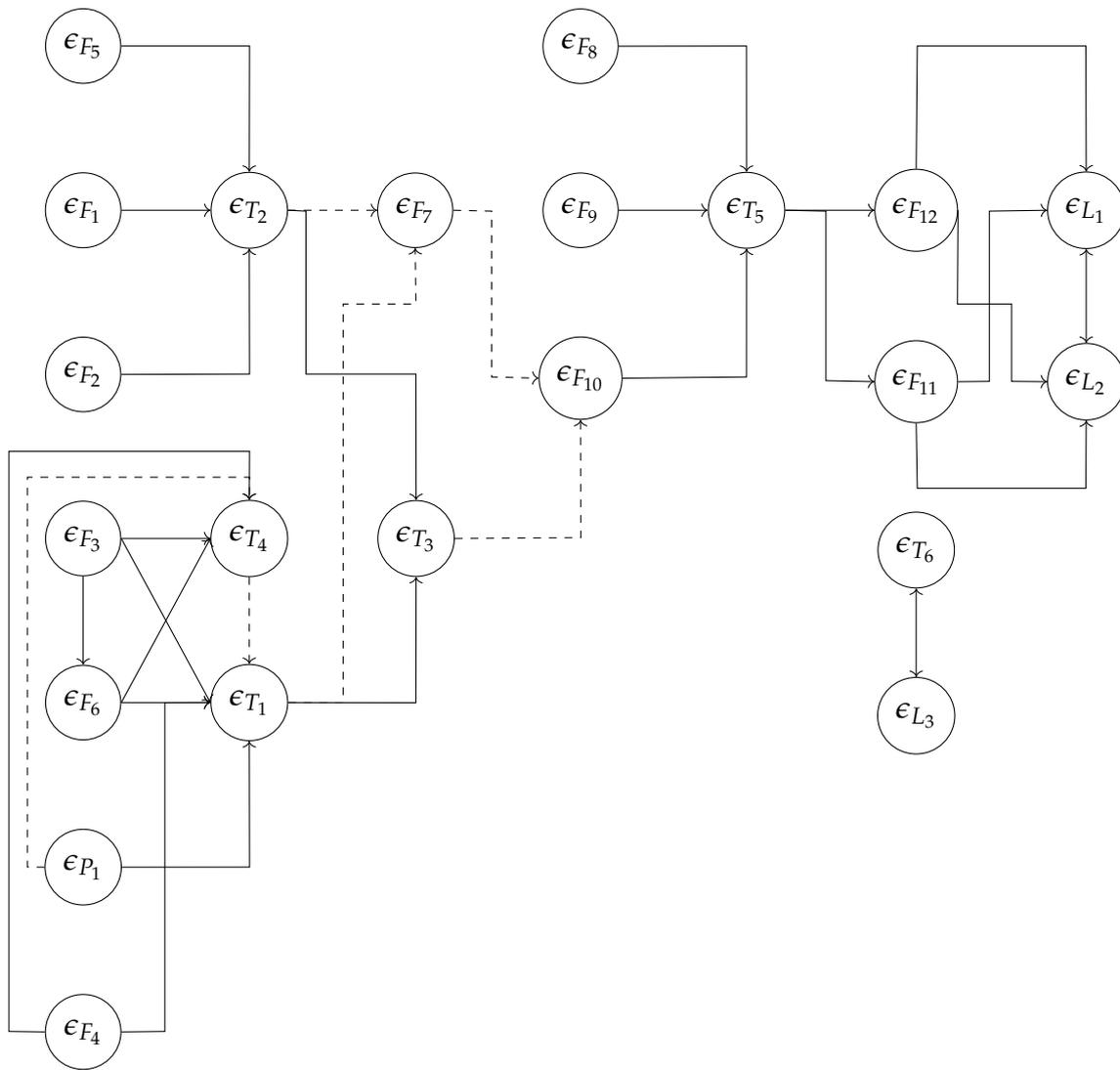


FIGURE 4.23: The actual causality diagram of the ammonia synthesis installation

Chapter 5

Implementation of the classical Transfer Entropy method

In this section, the relationships between the five control error variables of a simulation system are analyzed using the Transfer Entropy method in its basic version (based on the Gaussian distribution). Both simulation data, impacted by the Gaussian noise (see Figure 4.3) and by the Gaussian noise and Cauchy disturbance (see Figure 4.4) are assessed.

To better illustrate the fit of only the Gaussian distribution to the data, histograms for both datasets are shown in the following figures, representing individual control error variables. These histograms provide visual insights into the distribution of the data and help assess the adequacy of the Gaussian assumption. The x-axis represents the range of individual control error values, while the y-axis indicates the count of occurrence for each value range. Examining the histograms, we can observe the distribution pattern and assess whether it aligns with a Gaussian distribution. A good fit to the Gaussian distribution would typically show a symmetric bell-shaped curve centered around the mean value.

In typical scenarios involving data analysis, apart from constructing histograms, it is customary to employ normality tests such as the Kolmogorov-Smirnov or Lilliefors. These tests serve a vital purpose when there is a fundamental need to scrutinize whether a given dataset adheres to the fundamental assumptions of normality that underlie a multitude of statistical methodologies. However, the judicious application of normality tests hinges upon the intrinsic characteristics of the dataset at hand.

Control errors, as encountered in this case, are subject to the influence of Gaussian noise and Cauchy disturbance. Such influence unequivocally engenders deviations from the idealized normal distribution, imparting a non-normal quality to the observed error patterns. Consequently, the very nature of these errors diverges from Gaussian expectations.

The discerning choice to abstain from subjecting the dataset to normality tests is steered by an array of considerations, encompassing the wealth of prior knowledge pertaining to the simulation process. In instances where there exist compelling theoretical and practical rationales that firmly suggest the inappropriateness of a normal

distribution as a descriptor for the dataset, the application of normality tests stands to yield limited additional insights.

Figure 5.1 shows a histogram for control error ϵ_1 from the simulation system affected by Gaussian noise. Visual inspection shows that the fit is not perfect and can be improved with data processing techniques, or by changing the PDF type entirely.

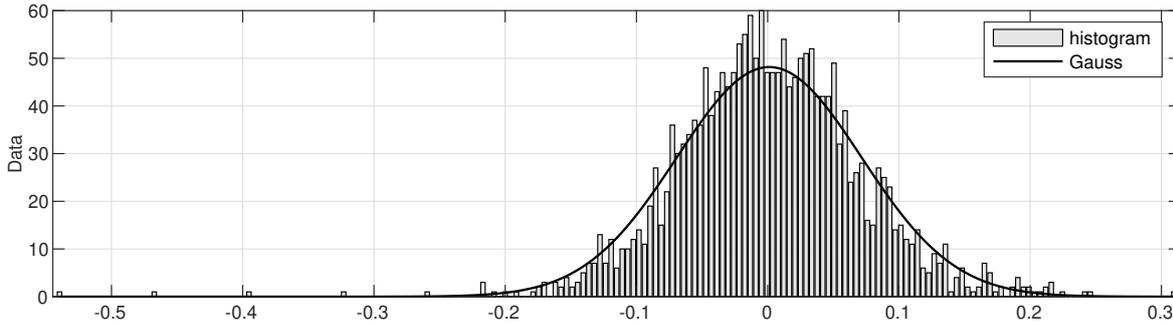


FIGURE 5.1: Control error ϵ_1 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting

The histogram for the control error variable ϵ_2 from the simulation system affected by Gaussian noise is presented in Figure 5.2. Upon visual inspection, it can be observed that the fit of the Gaussian distribution is comparatively better for ϵ_2 when compared to ϵ_1 . However, there is still room for improvement in the fitting of the Gaussian distribution for ϵ_2 .

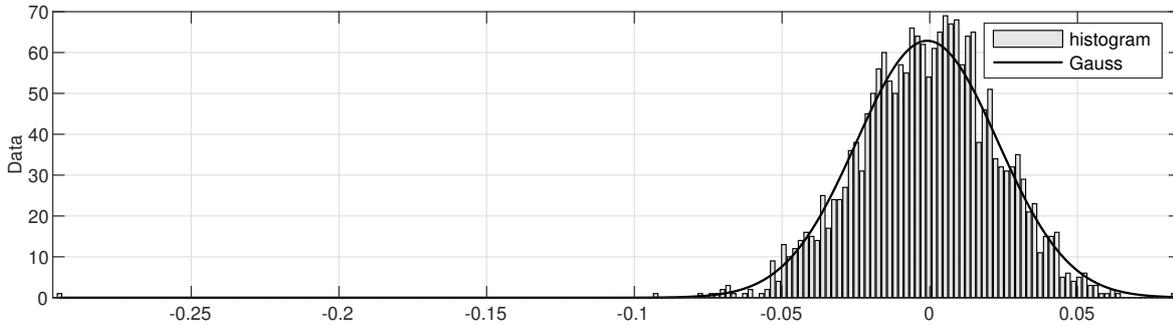


FIGURE 5.2: Control error ϵ_2 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting

Similar to the control error ϵ_1 , the fit of the Gaussian distribution for control error ϵ_3 is also inadequate. This observation is evident from the histogram shown in Figure 5.3. The distribution of ϵ_3 deviates noticeably from the expected Gaussian shape, indicating that the assumption of a pure Gaussian noise model might not accurately capture the underlying dynamics of this variable. Further analysis and modeling techniques may be required to improve the fit and gain a better understanding of the behavior of ϵ_3 in the simulation system.

Chapter 5. Implementation of the classical Transfer Entropy method

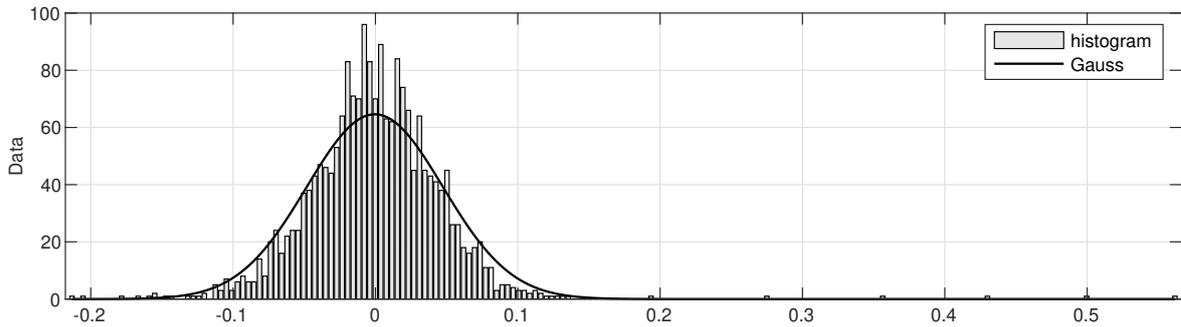


FIGURE 5.3: Control error ϵ_3 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting

The situation worsens when considering control error variables ϵ_4 and ϵ_5 . The application of a parametric Gaussian model fails to capture the dynamics of both control errors effectively. This is evident from the histograms presented in Figure 5.4 and Figure 5.5, where the distributions significantly deviate from the Gaussian shape.

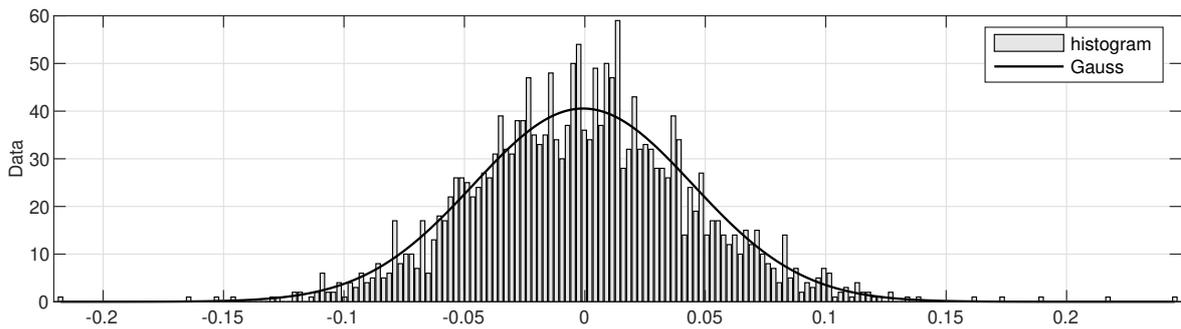


FIGURE 5.4: Control error ϵ_4 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting

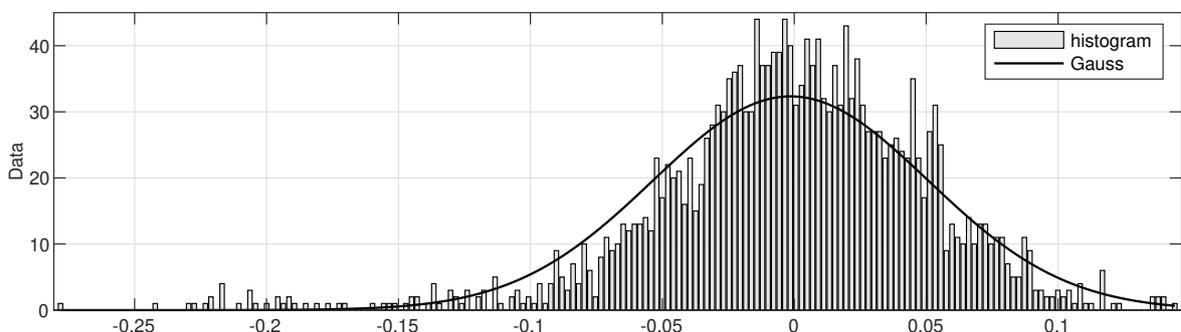


FIGURE 5.5: Control error ϵ_5 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting

The subsequent histograms provided in Figure 5.6, Figure 5.7, Figure 5.8, Figure 5.9, and Figure 5.10 depict the distribution of control error variables in the simulation system, when influenced by both Gaussian noise and Cauchy disturbance.

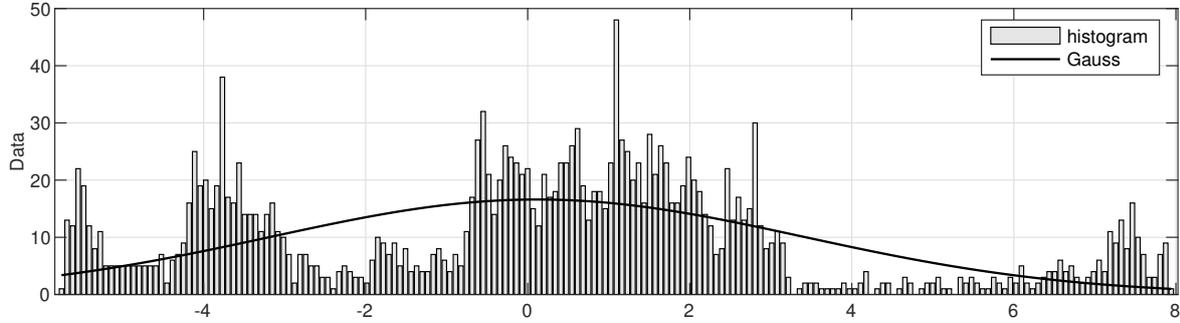


FIGURE 5.6: Control error ϵ_1 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting

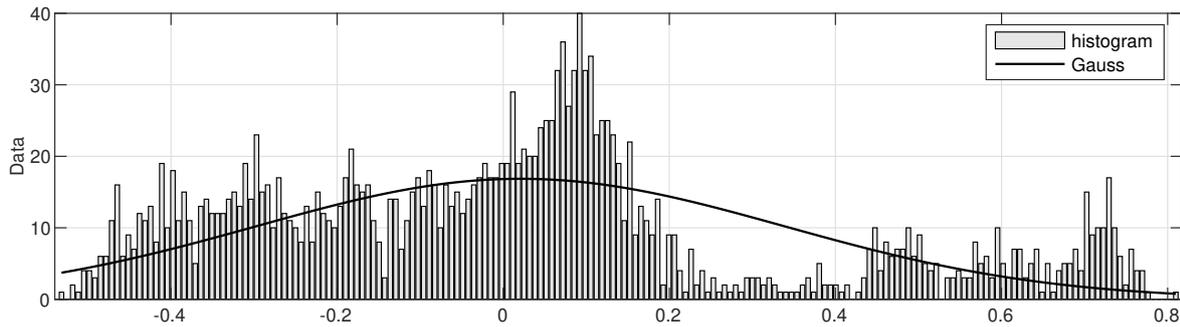


FIGURE 5.7: Control error ϵ_2 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting

By analyzing these histograms, we can gain insights into how the presence of the fat-tailed Cauchy disturbance affects the distribution of each control error variable. These histograms allow us to compare the resulting distributions with the histograms obtained when considering only Gaussian noise. Such a comparison can help us evaluate the impact of the Cauchy disturbance on the shape and characteristics of the control error variables. The inclusion of Cauchy disturbance in the dataset may lead to further deviations from the Gaussian model, as Cauchy distributions have heavier tails compared to Gaussian distributions. By examining these histograms, we can assess the extent to which the presence of Cauchy disturbance affects the fit of the Gaussian model and the overall distribution of the control error variables.

Fitting a Gaussian model to the data for control error variables ϵ_1 and ϵ_2 for the given dataset proves to be extremely challenging, if not impossible. The nature of

the data itself does not exhibit a normal distribution pattern, as evident from the histograms shown in Figure 5.6 and Figure 5.7. The lack of a normal distribution for these control errors indicates that relying on a Gaussian model would not accurately capture their underlying dynamics. It is crucial to acknowledge that attempting to force a Gaussian assumption onto such non-Gaussian data can lead to inaccurate analysis and misleading results.

When considering control error variables ϵ_3 and ϵ_4 in the simulation system, the situation appears to show some improvement compared to the previous cases. However, it is still far from meeting the expectations set by the Gaussian model. The histograms presented in Figure 5.8 and Figure 5.9 illustrate the distribution of the data for these control errors. While there are some similarities to a Gaussian distribution, there are noticeable deviations that prevent a perfect fit.

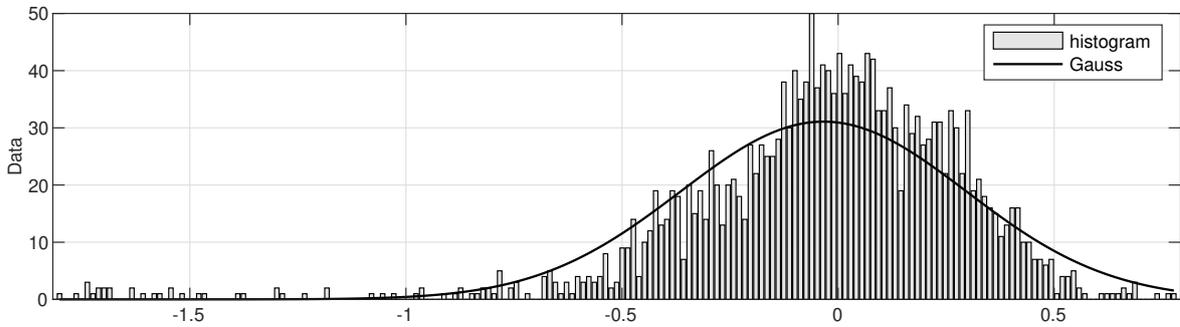


FIGURE 5.8: Control error ϵ_3 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting

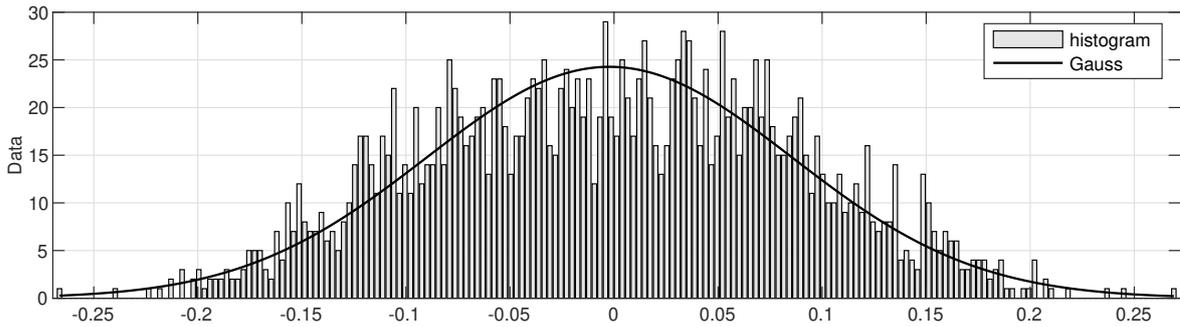


FIGURE 5.9: Control error ϵ_4 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting

Similar to the cases of control errors ϵ_1 and ϵ_2 , control error ϵ_5 exhibits a significant deviation from the Gaussian model. The nature of the data demonstrates a clear departure from a normal distribution pattern. Upon examination of the histogram shown in Figure 5.10, it is apparent that the distribution of ϵ_5 does not align with

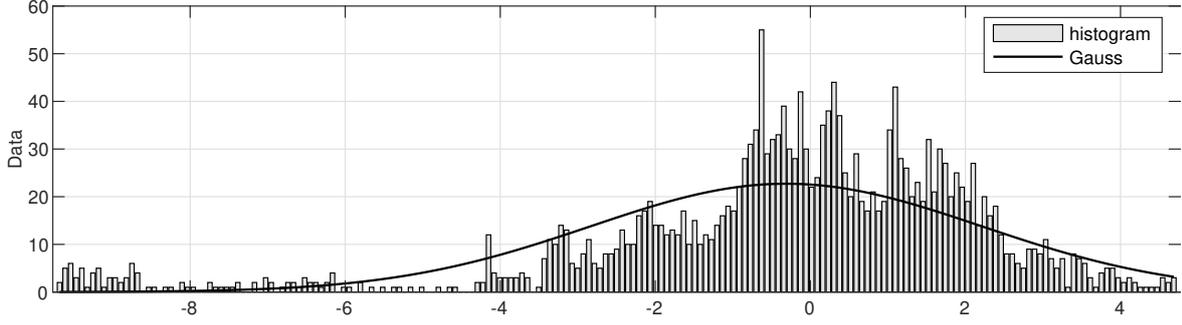


FIGURE 5.10: Control error ϵ_5 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting

the expectations set by the Gaussian model. The observed deviations highlight the presence of underlying dynamics or influence that render the Gaussian assumption inadequate for accurately representing the behavior of ϵ_5 .

Despite the challenges of fitting a Gaussian distribution to the data, it is still possible to apply the Transfer Entropy method to both datasets, including the one impacted by Gaussian noise and the one affected by both Gaussian noise and Cauchy disturbance. While the Gaussian assumption may not accurately represent the underlying dynamics of the control error variables, the Transfer Entropy method is not only limited to specific distributional assumptions. Therefore, according to the Eq. 3.14, the following parameters presented in Table 5.1 are defined for each dataset calculation.

TABLE 5.1: Parameters of Transfer Entropy function for the simulation data

Parameter	Relationship	Value	Description
τ	$\epsilon_1 \leftrightarrow \epsilon_2$	0	time delay between control errors
	$\epsilon_1 \leftrightarrow \epsilon_3$	9	time delay between control errors
	$\epsilon_1 \leftrightarrow \epsilon_4$	0	time delay between control errors
	$\epsilon_1 \leftrightarrow \epsilon_5$	102	time delay between control errors
	$\epsilon_2 \leftrightarrow \epsilon_3$	0	time delay between control errors
	$\epsilon_2 \leftrightarrow \epsilon_4$	0	time delay between control errors
	$\epsilon_2 \leftrightarrow \epsilon_5$	97	time delay between control errors
	$\epsilon_3 \leftrightarrow \epsilon_4$	1	time delay between control errors
	$\epsilon_3 \leftrightarrow \epsilon_5$	62	time delay between control errors
$\epsilon_4 \leftrightarrow \epsilon_5$	9	time delay between control errors	
N_p	-	20	value of 20 points along each dimension is set
C_{thumb}	-	1	rule of thumb

In situations where the time-series length is relatively short, a time delay parameter τ of 1 is often chosen when applying the Transfer Entropy method. This

5.1. Causality for the dataset with Gaussian noise

assumption is based on the notion that the maximum transfer of information typically occurs from the data point immediately preceding the target value in y . These parameter values, according to Table 5.1, are selected based on the specific characteristics of the dataset and the desired analysis goals.

Another important parameter to consider is N_p , which determines the number of points along each dimension used for estimating probabilities in the Transfer Entropy calculation. Sample tests conducted in the research indicate that varying N_p between 20, 30, 40, and 50 leads to negligible differences in calculated entropies. Therefore, a value of $N_p = 20$ is chosen as a practical compromise (as shown in Table 5.1).

It is worth noting that the parameter C_{thumb} set to 1 in this case, indicates that there are no modifications made to the default rule of thumb for selecting parameter values. These parameter choices and adjustments aim to optimize the Transfer Entropy method for the specific dataset, considering the limited time-series length and other practical considerations.

With a careful selection of these parameter values, we can ensure that the Transfer Entropy analysis effectively captures the information flow and dependencies between the control error variables, providing meaningful insights into their relationships in the simulation system.

5.1 Causality for the dataset with Gaussian noise

When analyzing the simulation data affected by Gaussian noise, it becomes challenging to present the results in the form of TE coefficients. The calculated values presented in Table 5.2 often result in "NaN" (Not a Number) or "Inf" (Infinity) entries, which are typically interpreted as zero in the context of causality analysis. This implies that no causality is observed between such variables.

TABLE 5.2: Calculated Transfer Entropy coefficients for the simulation data – Gaussian noise

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	Inf	NaN	NaN	NaN
ϵ_2	0.0762	NA	NaN	NaN	NaN
ϵ_3	NaN	NaN	NA	NaN	Inf
ϵ_4	NaN	0.0288	Inf	NA	NaN
ϵ_5	NaN	Inf	Inf	NaN	NA

The prevalence of "NaN" or "Inf" values in the results can be attributed to the Gaussian disturbance being too weak to adequately stimulate the simulation system. The impact of the Gaussian noise on the variables may not be significant enough to produce discernible causal relationships or information transfer captured by the Transfer Entropy method.

It is important to recognize the limitations of the Gaussian noise and its impact on the simulation system. The weak stimulation resulting from the Gaussian disturbance can lead to inconclusive or non-existent causal relationships, as reflected in the "NaN" or "Inf" values observed in the analysis.

To address this issue and to obtain more meaningful results, it may be necessary to consider alternative approaches such as increasing the intensity of the Gaussian disturbance or exploring different types of disturbances that can better stimulate the simulation system, such as adding the Cauchy disturbance.

5.2 Causality for the dataset with Gaussian noise and Cauchy disturbance

Results for a second case with simulation data impacted by Gaussian noise and Cauchy disturbance are presented in Table 5.3.

TABLE 5.3: Calculated Transfer Entropy coefficients for the simulation data – Gaussian noise and Cauchy disturbance

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.1143	0.0743	0.0608	0.0631
ϵ_2	0.0589	NA	0.0556	0.0490	0.0653
ϵ_3	0.0568	0.0475	NA	0.0699	0.0854
ϵ_4	0.0285	0.0374	0.0547	NA	0.0419
ϵ_5	0.0685	0.0694	0.0886	0.0646	NA

Applying a threshold value to the obtained entropy values in Transfer Entropy analysis is subjective and is not an objective approach. To facilitate interpretation, a practical approach is adopted, where the highest entropy value in each row is identified and visually highlighted in blue and bold font. This helps draw attention to the most prominent entropy value for each pair of variables.

The resulting causality graph for this specific case with comparison to the actual graph is depicted in Figure 5.11. Comparing this graph to the assumed causality diagram, it becomes apparent that there is little resemblance between them. This discrepancy suggests that the Transfer Entropy method might not effectively capture the expected causal relationships in the simulated data.

One possible reason for this poor performance could be attributed to the inherent characteristics of the data itself. It is observed that the variables exhibit continuous changes and display clear trends over time. Such characteristics, including the presence of sudden changes or discontinuities, can pose challenges for the Transfer Entropy analysis.

Investigations of real data (Falkowski and Domański, 2020) reveal that Transfer Entropy is not resilient to sudden changes in variable values. This lack of resilience

5.2. Causality for the dataset with Gaussian noise and Cauchy disturbance

indicates that the Transfer Entropy approach may be ineffective in accurately capturing the relationships and determining reliable coefficients in such scenarios.

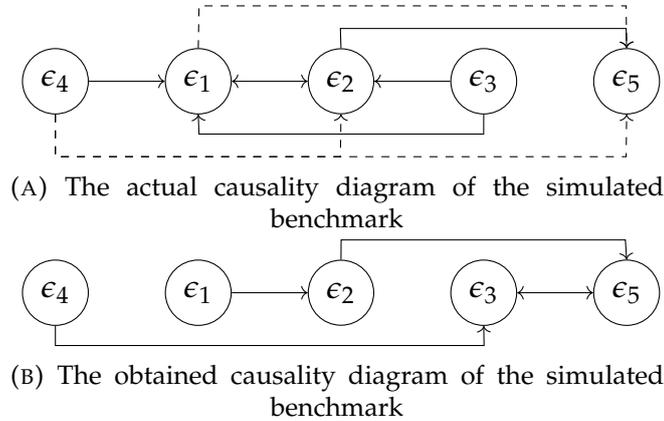


FIGURE 5.11: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance

Based on these findings, it is advised to take serious caution to the determined coefficient values obtained through the Transfer Entropy method. It is crucial to explore alternative analysis approaches or models that can better handle the specific characteristics of the data and provide more reliable results. The inadequate fit of the Gaussian model to these control errors has significant implications when using the Transfer Entropy method. Transfer Entropy relies on accurate modeling and assumptions about the underlying dynamics of the variables involved. In this case, the deviation from a Gaussian distribution shows that the Transfer Entropy results obtained using the Gaussian model are unreliable and misleading.

To overcome this limitation, alternative modeling approaches or non-parametric methods that can capture the non-Gaussian dynamics of data are required. The Transfer Entropy method can be applied with greater confidence by using more appropriate models, leading to more accurate and meaningful results in the analysis of the relationships between these control error variables in the simulation system.

Chapter 6

Impact of the PDF on the Transfer Entropy method

In this chapter, the calculation results based on the Transfer Entropy approach are presented, with a focus on comparing the outcomes obtained using parametric and non-parametric statistical models. The aim is to assess the impact of the choice of probability density function (PDF) determination on the simulation system. The results are discussed in two sections: Section 6.1 highlights the findings using parametric statistical models, while Section 6.2 delves into the outcomes obtained with non-parametric statistical models.

6.1 Parametric statistical models

The causality analysis using the Transfer Entropy approach continues with the determination of parametric probability distributions, adjusted for each control error separately. This applies to data generated both with Gaussian noise as well as for the Gaussian noise and Cauchy disturbance. The fittings of Cauchy, α -stable, Laplace, Huber, and t Location-scale distributions for the dataset with Gaussian noise are presented in Figure 6.1, Figure 6.2, Figure 6.3, Figure 6.4, and Figure 6.5, for each ϵ_i respectively.

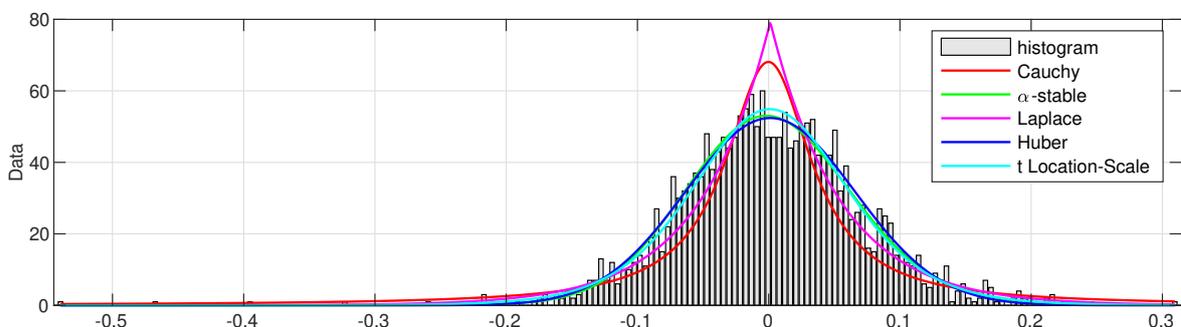


FIGURE 6.1: Control error ϵ_1 histogram for the simulation data including Gaussian noise with selected PDF fitting

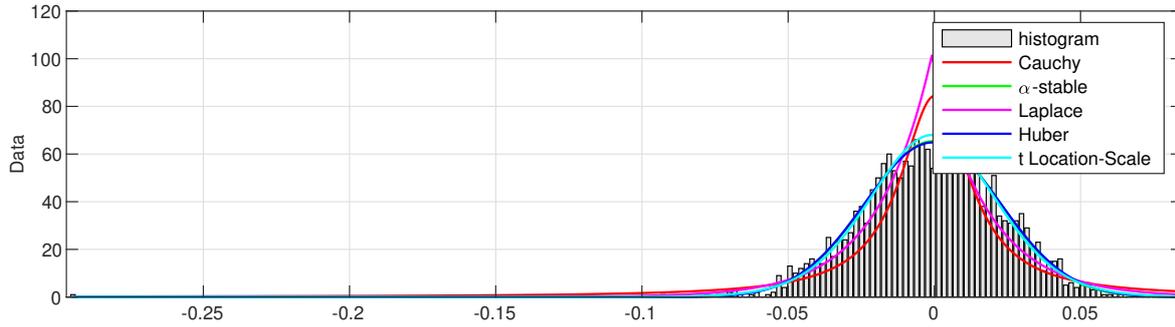


FIGURE 6.2: Control error ϵ_2 histogram for the simulation data including Gaussian noise with selected PDF fitting

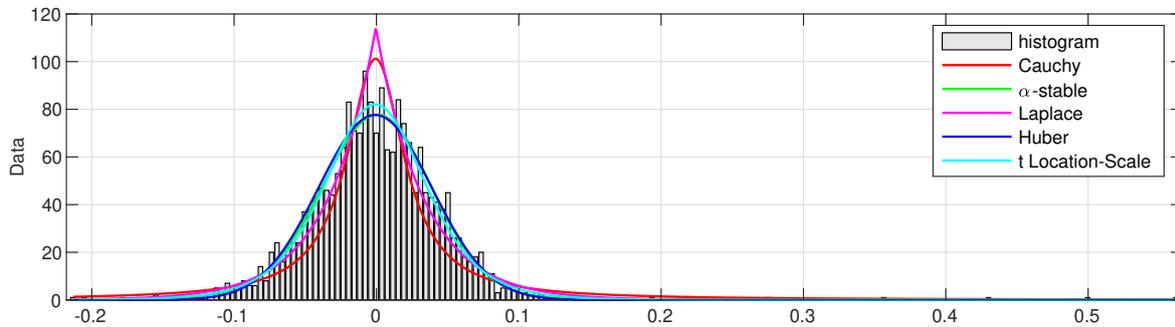


FIGURE 6.3: Control error ϵ_3 histogram for the simulation data including Gaussian noise with selected PDF fitting

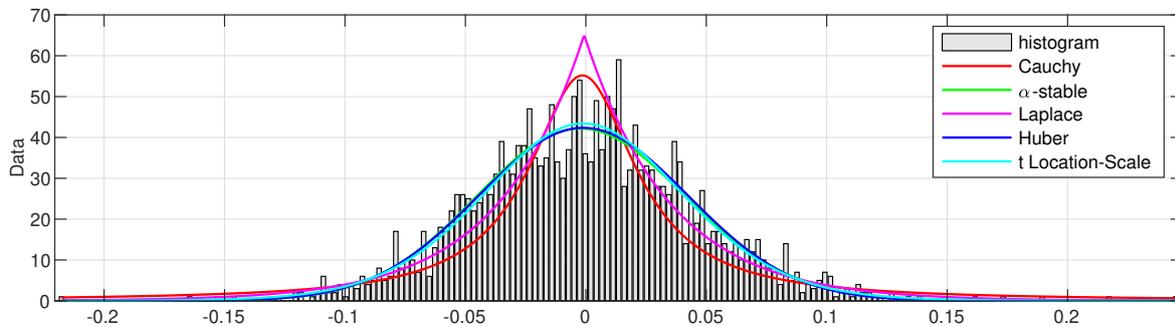


FIGURE 6.4: Control error ϵ_4 histogram for the simulation data including Gaussian noise with selected PDF fitting

Upon visual inspection, it becomes evident that the Cauchy and Laplace probability distributions do not adequately fit the observed data. The patterns and characteristics exhibited by the observed data deviate significantly from those expected from the Cauchy and Laplace distributions. The lack of fit suggests that these distributions may not be appropriate for accurately modeling the data. However, a slightly better fit is observed when considering alternative distributions such as the α -stable, Huber, and t Location scale. Although they still do not provide a perfect fit

6.1. Parametric statistical models

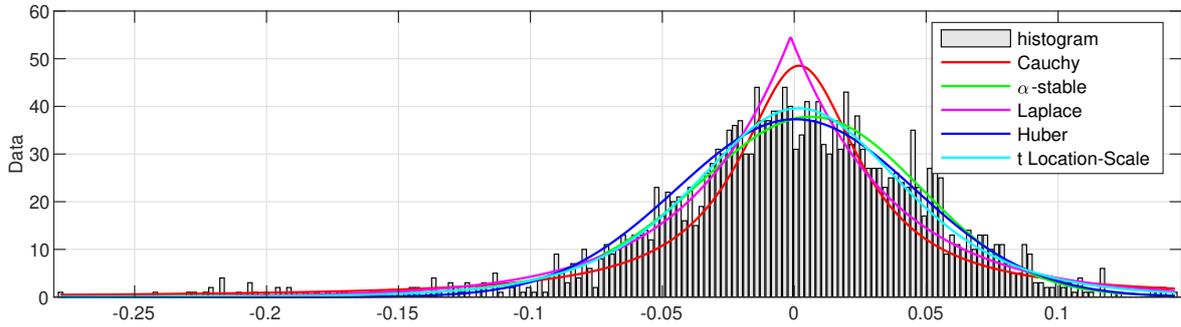


FIGURE 6.5: Control error ϵ_5 histogram for the simulation data including Gaussian noise with selected PDF fitting

but demonstrate a closer alignment with the observed data.

For the dataset with Gaussian noise and Cauchy disturbance, fittings of Cauchy, α -stable, Laplace, Huber, and t Location-scale distributions are presented in Figure 6.6, Figure 6.7, Figure 6.8, Figure 6.9, and Figure 6.10, for each ϵ_i respectively.

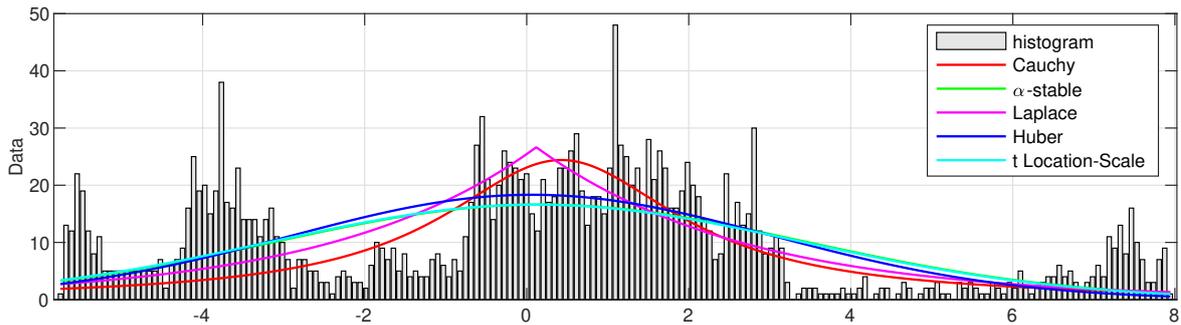


FIGURE 6.6: Control error ϵ_1 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting

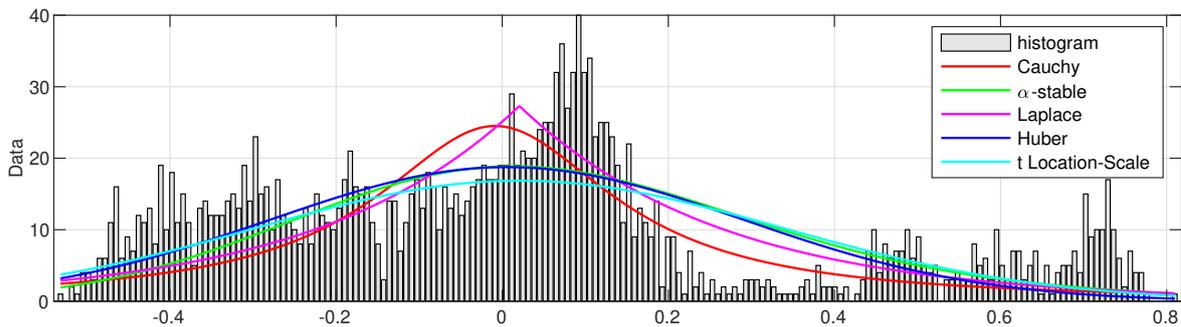


FIGURE 6.7: Control error ϵ_2 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting

Upon closer examination, a significant portion of the data does not exhibit sufficient normality. This lack of normality is particularly pronounced in the case of

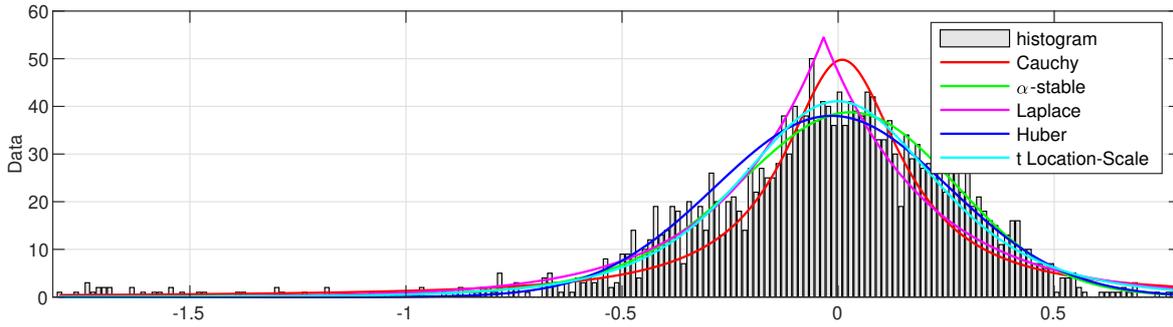


FIGURE 6.8: Control error ϵ_3 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting

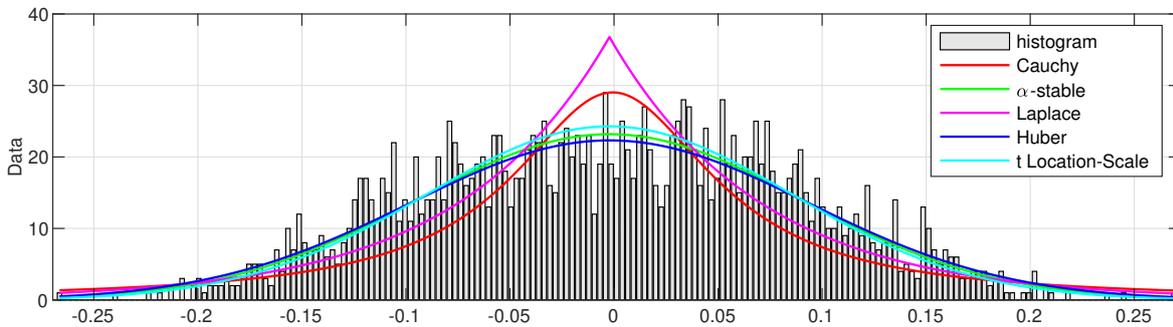


FIGURE 6.9: Control error ϵ_4 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting

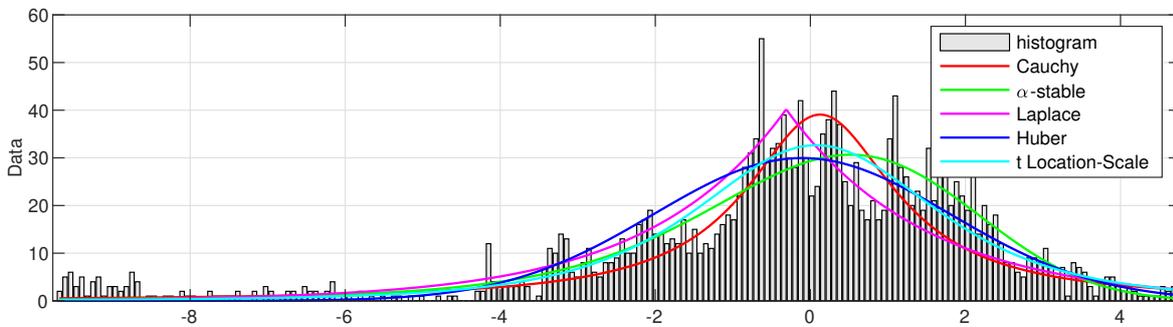


FIGURE 6.10: Control error ϵ_5 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting

the control errors ϵ_1 and ϵ_2 . The characteristics of these errors do not conform to any of the given distributions considered for fitting. The inability to fit the data to any of the provided distributions suggests that the underlying nature of the data is more complex and does not adhere to the assumptions. In such cases, alternative approaches may need to be considered. In Chapter 3.5, non-parametric methods are explored to address the non-normality and better capture the true behavior of such data.

6.1. Parametric statistical models

Enhancing the goodness of fit between the chosen parametric probability distributions and the observed data is crucial for accurate modeling, reliable predictions, and gaining meaningful insights into the underlying phenomena. Nevertheless, the assessment of the quality of fit of the probability distributions selected for analysis can not be only a visual evaluation. To objectively determine which PDF fits the control error histogram the best, a fit indicator is used. The fitting indexes are calculated as mean square errors between the empirical histogram heights and the value of the fitted probabilistic density function evaluated at the middle of the respective histogram bin. The results for the dataset with Gaussian noise are presented in Table 6.1 and for the dataset with Gaussian noise and Cauchy disturbance are presented in Table 6.2. The lowest values of the fit indicator (which is considered the best fitting) for each ϵ_i are marked in blue, and bold.

TABLE 6.1: Fit indicators for the dataset with Gaussian noise

	Cauchy	α -stable	Laplace	Huber	t Location-scale
ϵ_1	3.3440	1.4925	2.4805	1.5098	1.5333
ϵ_2	3.6252	1.3049	2.7709	1.3007	1.4069
ϵ_3	3.3966	1.4344	2.2947	1.5182	1.5246
ϵ_4	3.5771	1.9684	2.8030	1.9850	1.9877
ϵ_5	3.1062	1.9417	2.5749	1.9473	1.8361

TABLE 6.2: Fit indicators for the dataset with Gaussian noise and Cauchy disturbance

	Cauchy	α -stable	Laplace	Huber	t Location-scale
ϵ_1	5.1540	5.7679	5.5035	5.6962	5.7589
ϵ_2	5.0164	5.0727	4.9170	4.9768	4.9432
ϵ_3	3.1914	1.9279	2.9657	2.1143	2.1835
ϵ_4	3.9075	2.2738	3.6509	2.3228	2.3033
ϵ_5	3.5362	3.1455	3.8385	3.4425	3.1752

Table 6.1 shows that for the dataset with Gaussian noise, α -stable distribution appears to be best fitted in case of control errors ϵ_1 , ϵ_3 , and ϵ_4 . For the control error ϵ_2 it is robust distribution and for the ϵ_5 , the t location-scale. In Table 6.2, the lowest value of the fit indicator is calculated for the α -stable distribution for control errors ϵ_3 , ϵ_4 , and ϵ_5 . Significantly different nature of trends of the control errors ϵ_1 and ϵ_2 , clearly affected the fit indicators calculations. In both cases, the values of the indicator are higher than those determined for control errors ϵ_3 , ϵ_4 , and ϵ_5 . Still, the

best fit for ϵ_1 is characterized by the Cauchy distribution, and for the ϵ_2 , the Laplace distribution.

The general conclusion for both datasets is that the α -stable distribution in most cases has the best fit for considered ϵ_i . Therefore, it is decided that concerning the classical Transfer Entropy approach, the determined TE coefficients are based on α -stable distribution. The results for the dataset with Gaussian noise and the dataset with Gaussian noise and Cauchy disturbance are presented in Section 6.1.1 and Section 6.1.2 respectively. The parameters of the Transfer Entropy approach remain the same (see Table 5.1).

6.1.1 Causality for the dataset with Gaussian noise

In this section, the simulation data impacted by the Gaussian noise (see Figure 4.3) is analyzed. The relationships between five control errors of a simulation system are designated using the Transfer Entropy approach based on Gaussian distribution and α -stable distribution.

It is shown that results for simulation data impacted by Gaussian noise for the classic Transfer approach are problematic to be presented in the form of a table (see Section 5.1). When using the Transfer Entropy method based on α -stable distribution, the results turn out to be satisfactory. The TE coefficients are obtained, included in Table 6.3.

TABLE 6.3: Calculated Transfer Entropy coefficients based on α -stable distribution for the dataset with Gaussian noise

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.1284	0.0729	0.0725	0.1016
ϵ_2	0.1499	NA	0.0508	0.0387	0.0511
ϵ_3	0.0450	0.0531	NA	0.0322	0.0401
ϵ_4	0.0435	0.0371	0.0349	NA	0.0399
ϵ_5	0.0430	0.0518	0.0409	0.0433	NA

Since there is no objective method for applying a threshold value for obtained entropy values, the highest entropy value in each row is marked in blue, and bold. The above assumption enforces to show only the direct relationships but still, the causality graph can be drawn. It is given in Figure 6.11b.

Based on the analysis of the results, it can be concluded that the change in the probability distribution has a positive impact on the quality of the Transfer Entropy (TE) method. This suggests that by considering a different probability distribution, the TE method is better able to capture and represent the relationships and causal dependencies within the data.

6.1. Parametric statistical models

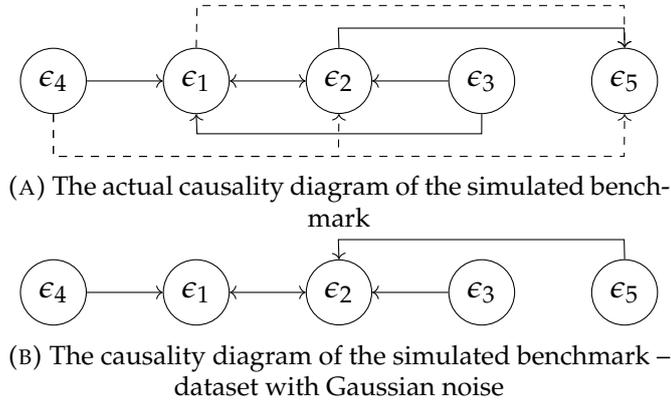


FIGURE 6.11: Comparison of the actual and obtained causality graphs - data impacted by Gaussian noise

Scrutinizing the causality graph obtained from the analysis and comparing it to the assumed causal relationships depicted in Figure 6.11a, noticeable discrepancies emerge in the directional relationships between the control errors ϵ_2 and ϵ_5 . Additionally, an absence of correlation is observed between ϵ_3 and ϵ_1 , contrary to the assumed causality. These inconsistencies can be attributed to the underlying assumption used in the analysis, which relies on determining the relationship between control errors based on the highest Transfer Entropy (TE) coefficient within a given row. It is shown that the reliance on the highest TE coefficient in a row to establish causality can lead to misinterpretations, especially in a given scenario, where the relationship between variables is more complex or when multiple causal factors are at play.

The discrepancy in the direction of the relationship between ϵ_1 and ϵ_5 suggests that further investigation and analysis may be necessary to better understand the underlying dynamics and causal mechanisms in this particular relationship. It could be due to various factors such as non-linear dependencies or the influence of external factors, such as outliers, not accounted for in the analysis. It highlights the importance of critically evaluating and interpreting the results of this analysis method. In the case of real-world applications, it is crucial to consider the limitations and potential sources of error, such as model assumptions or data quality issues, that could impact the accuracy of the identified causal relationships.

6.1.2 Causality for the dataset with Gaussian noise and Cauchy disturbance

This section focuses on the analysis of simulation data that has been influenced by Gaussian noise and Cauchy disturbance, as illustrated in Figure 4.4. To examine the causal relationships within this data, the Transfer Entropy (TE) approach utilizing the α -stable distribution is employed. The calculated entropies in this case are documented in Table 6.4.

TABLE 6.4: Calculated Transfer Entropy coefficients based on α -stable distribution for the dataset with Gaussian noise and Cauchy disturbance

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.1154	0.0729	0.0599	0.0601
ϵ_2	0.0561	NA	0.0540	0.0491	0.0666
ϵ_3	0.0568	0.0475	NA	0.0699	0.0854
ϵ_4	0.0772	0.0284	0.0521	NA	0.0380
ϵ_5	0.0643	0.0695	0.0888	0.0623	NA

Furthermore, the resulting causality graph associated with Table 6.4 is sketched in Figure 6.12b.

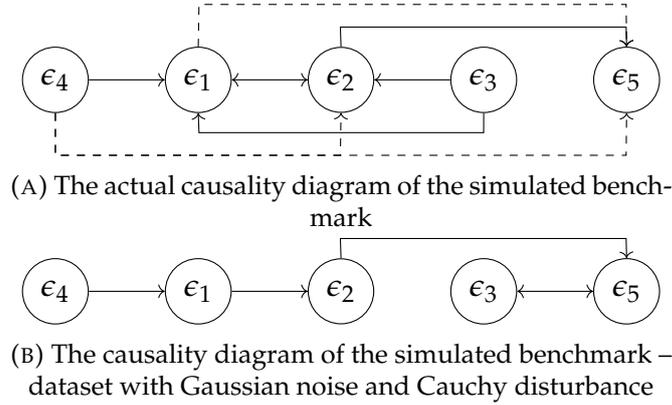


FIGURE 6.12: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance

The resulting causality relationships obtained from the Transfer Entropy calculations are found to be only partially correct. Precise indications can be made for the relationships between control errors ϵ_4 and ϵ_1 , as well as ϵ_2 and ϵ_5 . However, other relationships inferred from the analysis are misleading and do not align with the assumed causality graph presented in Figure 6.12a.

These inconsistencies raise several questions regarding the obtained results. In particular, the effectiveness of changing the probability density function from a Gaussian distribution to an α -stable distribution for a dataset with Gaussian noise is questionable. It is evident that we deal with data of a different nature, requiring deeper analysis and understanding. The use of non-parametric statistical models presents another clue or indication for the analysis. Unlike parametric models, the non-parametric ones make fewer assumptions about the underlying data distribution or functional form (see Section 3.3.1, Section 3.3.2, and Section 3.3.3). Instead, they rely on ranking, ordering, or other distribution-free techniques to analyze the data.

6.2 Non-parametric statistical models

Within this section, there is applied a set of non-parametric statistical models, named the Darbellay-Vajda algorithm, Fixed Bins, and Kernel Density Estimation, in conjunction with the Transfer Entropy approach. To ensure the accuracy of the analysis, the methodologies outlined in Section 3.3.1, Section 3.3.2, and Section 3.3.3 are followed. For the Fixed Bins approach, the parameter Q is checked for each pair of control errors separately. Similarly, for the Kernel Density Estimation approach, the parameter λ is determined. In the majority of cases, the Q and λ parameters exhibit repeatability. For calculations, there are adopted fixed values of $Q = 8$ and $\lambda = 1$. These parameters remain consistent throughout the analysis, ensuring reproducibility and comparability across given datasets.

6.2.1 Causality for the dataset with Gaussian noise

The initial findings of Transfer Entropy calculations using the Darbellay-Vajda algorithm are presented in Table 6.5. Remarkably, despite employing a straightforward estimation of time lags for the specified pairs of control errors, the results demonstrate a high level of accuracy. These outcomes highlight the effectiveness of the Darbellay-Vajda algorithm in capturing the underlying causal relationships within the data.

TABLE 6.5: Calculated Transfer Entropy coefficients based on Darbellay-Vajda algorithm for the dataset with Gaussian noise

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.0665	0	0.0726	0.0412
ϵ_2	0.1411	NA	0	0.0529	0.0193
ϵ_3	0.0899	0	NA	0.0057	0.0251
ϵ_4	0.1803	0.0439	0	NA	0.0173
ϵ_5	0.0252	0.0518	0	0	NA

It becomes evident upon a careful analysis of the causality graphs depicted in Figure 6.13 that the majority of the correct relationships among the control errors are indeed preserved. However, it is noteworthy that the direction of causality between ϵ_2 and ϵ_5 is inaccurately determined in this case. This discrepancy suggests the existence of hidden factors or complex dynamics influencing the causal interactions between these variables.

Another limitation arises when a particular control error, such as ϵ_3 , exhibits correlations with multiple errors, such as ϵ_1 and ϵ_2 . The current approach, which selects the highest Transfer Entropy coefficient value for a given set of relationships, fails to capture the full extent of the causal dependencies within the system. As a result, potential additional relationships among the variables may go unnoticed.

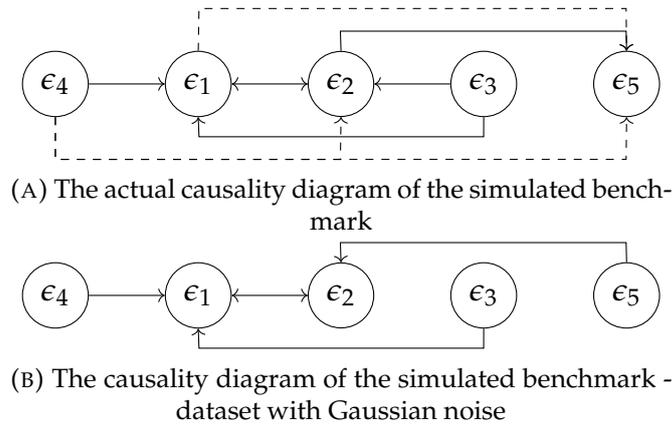


FIGURE 6.13: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise

The subsequent set of results, presented in Table 6.6, pertains to the application of the Transfer Entropy (TE) method utilizing the Fixed Bins algorithm. As described in Section 3.3.2, this methodology involves discretizing the dataset, partitioning it into fixed bins, and subsequently estimating the TE values based on this discretization approach. The table showcases the calculated TE values for the specified pairs of control errors, enabling the examination of the causal relationships within the dataset affected by Gaussian noise. By quantifying the TE values, valuable insights are gained regarding the strength and direction of the causal influences between the control errors.

TABLE 6.6: Calculated Transfer Entropy coefficients based on Fixed Bins algorithm for the dataset with Gaussian noise

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.2039	0.1772	0.1659	0.1773
ϵ_2	0.1716	NA	0.1612	0.1533	0.1438
ϵ_3	0.1427	0.1416	NA	0.1713	0.1786
ϵ_4	0.3216	0.1619	0.1734	NA	0.1634
ϵ_5	0.1438	0.1567	0.1546	0.1514	NA

Based on the results presented in Table 6.6, the causality graph for this particular case, along with the actual graph for comparison, is depicted in Figure 6.14. Notably, in the case of the Fixed Bins method, a Transfer Entropy coefficient greater than zero is obtained for each of the relationships in contrast to the results obtained using the Darbellay-Vajda (DV) algorithm. The Fixed Bins method partially succeeded in identifying the relationships between the control errors. This is evident, for instance, in the presence of a direct relationship between the ϵ_3 and ϵ_5 variables. However, it should be noted that the identified relationships are not entirely accurate, and

6.2. Non-parametric statistical models

there exist indirect influences between each of the errors. Additionally, studying the causality of the control error ϵ_5 poses challenges. Theoretically, each of the variables has an indirect relationship with ϵ_5 , as indicated by the TE coefficient values in the table. However, it is expected that the values of these coefficients would be orders of magnitude lower compared to the preceding control errors. These observations highlight the complexities involved in capturing the precise dependencies and interactions among the control errors, emphasizing the need for further refinement and exploration of alternative methodologies.

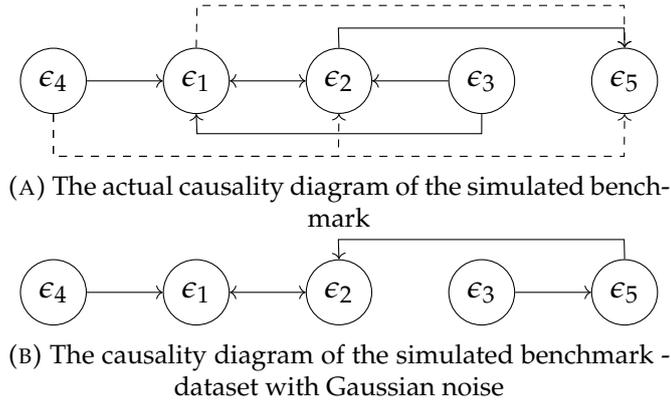


FIGURE 6.14: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise

The last non-parametric method of choice in this study is Kernel Density Estimation. The Transfer Entropy coefficients obtained using the KDE algorithm are presented in Table 6.7. This methodology involves estimating the TE values by constructing kernel density estimates for the probability distributions of the variables involved. The table provides a comprehensive overview of the TE coefficients, enabling the assessment of causal relationships within the dataset with Gaussian noise.

TABLE 6.7: Calculated Transfer Entropy coefficients based on Kernel Density Estimation method for the dataset with Gaussian noise

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.0726	0	0.0665	0.0412
ϵ_2	0.1411	NA	0	0.0529	0.0193
ϵ_3	0.0899	0	NA	0.0057	0.0251
ϵ_4	0.1803	0.0439	0	NA	0.0173
ϵ_5	0.0252	0.0518	0	0	NA

The findings derived from the analysis reported in Table 6.7 have been utilized to construct the causality graph specific to this case. This graph is visually compared to the actual causality graph, and both are illustrated in Figure 6.15.

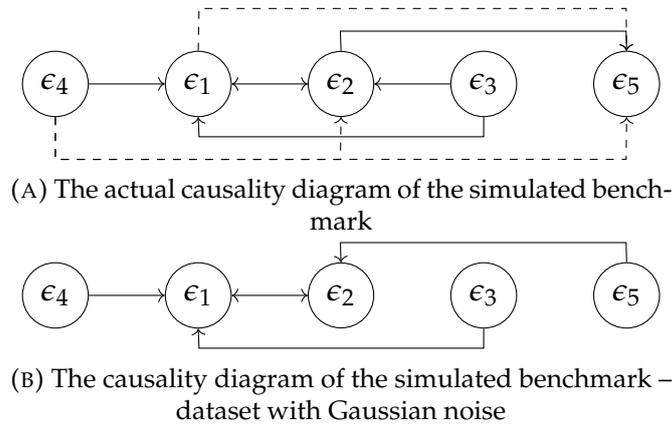


FIGURE 6.15: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise

Similar to the DV and Fixed bins case, the implementation of the Transfer Entropy method with the Kernel Density Estimation approach successfully reveals the relationships between the analyzed control errors. Once again, we encounter the issue of incorrect directionality in the relationship between ϵ_2 and ϵ_5 . In other instances, the obtained results are accurate.

The fundamental assumption guiding the determination of dependencies based solely on the largest TE coefficient leads to the failure in confirmation of the dependence between ϵ_3 and ϵ_2 . This particular assumption constitutes a significant limitation in the current approach. It is evident that this flaw will also affect the results obtained from the dataset with Gaussian noise and Cauchy disturbance, as discussed in Section 6.2.2.

To mitigate this problem, alternative rules to identify relationships could be explored. For instance, considering multiple significant TE coefficients or employing statistical tests to validate the causality relationships could enhance the reliability of the results.

6.2.2 Causality for the dataset with Gaussian noise and Cauchy disturbance

In this section, a comparative study using three distinct non-parametric statistical models such as Darbellay-Vajda, Fixed bins, and Kernel Density Estimation is conducted. The study focuses on the dataset with Gaussian noise and Cauchy disturbance. The preliminary results of Transfer Entropy calculations utilizing the Darbellay-Vajda algorithm are displayed in Table 6.8.

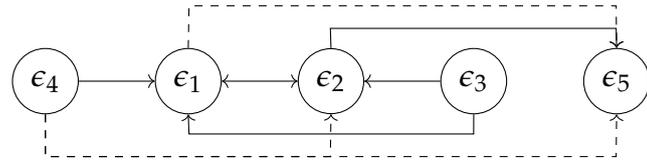
For each of the examined control errors, denoted as ϵ_i , the Transfer Entropy calculations yielded non-infinite (Inf) and non-not-a-number (NaN) coefficients, all of which are greater than zero. This observation indicates that meaningful information flow and directed causal interactions exist between the variables.

6.2. Non-parametric statistical models

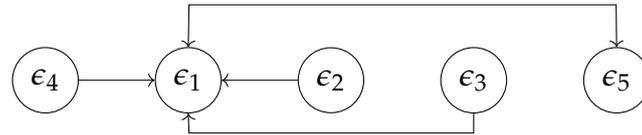
TABLE 6.8: Calculated Transfer Entropy coefficients based on Darbellay-Vajda algorithm for the dataset with Gaussian noise and Cauchy disturbance

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.6016	0.4445	0.1210	0.7697
ϵ_2	0.6458	NA	0.2475	0.1044	0.4532
ϵ_3	0.5236	0.2581	NA	0.0984	0.4128
ϵ_4	0.2571	0.2374	0.0989	NA	0.1808
ϵ_5	0.8251	0.3214	0.3061	0.0860	NA

Based on these obtained TE coefficients, a causality graph is constructed, representing the causal relationships and information transfer patterns between the control errors. Each edge in the causality graph corresponds to a significant TE coefficient, signifying the direction of causality from one control error to another. It is shown in Figure 6.16b.



(A) The actual causality diagram of the simulated benchmark



(B) The causality diagram of the simulated benchmark – dataset with Gaussian noise and Cauchy disturbance

FIGURE 6.16: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance

The examination of the causality graph in this case has brought to light an unexpected and unprecedented relationship between the control errors ϵ_1 and ϵ_5 . Upon comparing this graph to the actual causality graph illustrated in Figure 6.16a, it is evident that the identified relationship between ϵ_1 and ϵ_5 is solely indirect, precisely in direction $\epsilon_1 \rightarrow \epsilon_5$.

The emergence of this unique causal relationship between the control errors raises intriguing questions regarding the underlying dynamics and behavior of the given dataset. The presence of this causal relationship between ϵ_1 and ϵ_5 resulting from calculations suggests the possibility of mediating factors or unaccounted confounding variables influencing this relationship.

Despite notable improvements in the quality of the results compared to previous analyses, the current findings still exhibit inconsistencies with the actual causality graph. These discrepancies underscore the complexity inherent in the causal interactions within the dataset and underscore the need for more sophisticated methodologies.

The next notable non-parametric statistical model utilized in conjunction with the Transfer Entropy approach is the Fixed bins algorithm. The outcomes of this analysis are presented in Table 6.9.

TABLE 6.9: Calculated Transfer Entropy coefficients based on Fixed Bins algorithm for the dataset with Gaussian noise and Cauchy disturbance

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.2041	0.1820	0.1804	0.1908
ϵ_2	0.0885	NA	0.1615	0.1632	0.1830
ϵ_3	0.1234	0.0874	NA	0.1673	0.1829
ϵ_4	0.0768	0.1042	0.1288	NA	0.1172
ϵ_5	0.1301	0.1013	0.1642	0.1493	NA

According to acquired Transfer Entropy coefficients using the Fixed bins algorithm, the causality graph is visualized in Figure 6.17b.

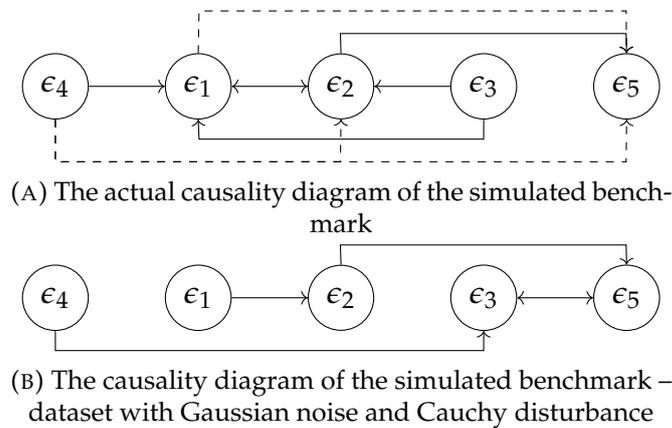


FIGURE 6.17: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance

For the dataset with Gaussian noise and Cauchy disturbance, the chosen approach encounters significant challenges and ultimately fails to accurately represent the underlying causal relationships. The resulting causality graph exhibits stark disparities when compared to the actual causality graph depicted in Figure 6.17a. The evident reason for this discrepancy lies in the unique nature of the analyzed data, which is heavily influenced by the presence of Cauchy disturbance.

6.2. Non-parametric statistical models

The pronounced impact of the Cauchy disturbance on the dataset poses considerable complexity in unraveling the true causal dependencies among control errors. The Fixed bins algorithm, which demonstrated relatively favorable performance in the absence of Cauchy disturbance, appears to struggle when confronted with this particular type of noise. Although it partially accomplishes its task by mapping some relationships, the presence of Cauchy disturbance evidently introduces significant challenges, leading to deviations from the actual causality graph.

This outcome serves as valuable feedback, highlighting the intricacies and limitations of the Fixed bins algorithm in the context of datasets influenced by diverse interference patterns. It underscores the crucial relationship between algorithm efficiency and the underlying characteristics of the analyzed data.

The final approach considered in this section involves the application of the Transfer Entropy approach utilizing the Kernel Density Estimation algorithm. Unlike the dataset with Gaussian noise, where several relationships resulted in TE coefficients equal to zero, the current analysis yielded non-zero TE coefficients for each of the examined control error pairs. The results are presented in Table 6.10.

TABLE 6.10: Calculated Transfer Entropy coefficients based on Kernel Density Estimation method for the dataset with Gaussian noise and Cauchy disturbance

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.1106	0.0320	0.0586	0.0294
ϵ_2	0.0605	NA	0.0256	0.0448	0.0308
ϵ_3	0.0526	0.0485	NA	0.0589	0.0412
ϵ_4	0.0243	0.0293	0.0224	NA	0.0152
ϵ_5	0.0520	0.0664	0.0308	0.0682	NA

The prominent positioning of the Transfer Entropy coefficient with the highest values in Table 6.10 unambiguously suggests that the identified associations between control errors in data affected by Gaussian noise and Cauchy disturbance are erroneous. Nevertheless, a causal graph can still be constructed, as depicted in Figure 6.18b.

Through a comprehensive comparison of the causality graph obtained using the Transfer Entropy method with a non-parametric statistical model in the form of Kernel Density Estimation, with the actual causality graph, it is evident that the KDE-based TE approach fails to capture the true causal relationships. This failure extends to both direct causal connections and even indirect relationships, which are denoted by dashed lines in the graph (see Figure 6.18a).

Attempts to adjust the methodology for this specific type of data are also unsuccessful. This included fine-tuning the method by altering the time lags or adjusting the parameter N , which represents the number of points along each dimension used

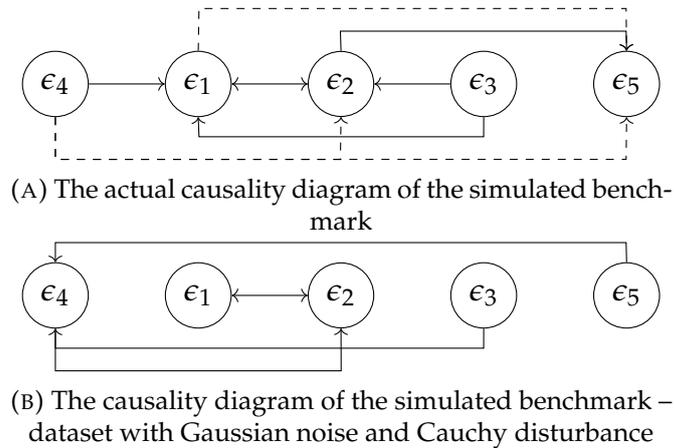


FIGURE 6.18: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance

in probability calculations. Despite these adjustments, the results remained perplexing and do not align with the expected causal relationships.

Interestingly, when the KDE-based TE method is applied to a dataset influenced solely by Gaussian noise, it showed promise and exhibited more favorable outcomes. However, when confronted with the additional factor of Cauchy disturbance, the approach proved to lack robustness and became inadequate in revealing the true causal patterns.

The observed discrepancy highlights the necessity for the implementation of advanced and sophisticated data analysis methods that can effectively handle the complexities arising from various types of noise and disturbances. In doing so, these methods should also enhance the efficacy of the Transfer Entropy approach in capturing accurate causal relationships. Rather than solely concentrating on developing more resilient causal inference techniques, future research should also emphasize the investigation of data properties, including the identification of outliers, as a means to address the challenges posed by this diverse dataset.

Chapter 7

Impact of outliers on the Transfer Entropy method

Understanding the impact of outliers and their locations in the dataset can significantly contribute to refining the causal inference process. Identifying and appropriately handling outliers can lead to a more accurate and reliable determination of causal relationships, even in the presence of complex noise patterns like Cauchy disturbance.

As a result, the future direction of research pursues a comprehensive approach that combines advanced outliers detection methods, ESD (see Section 3.5.1), IQR (see Section 3.5.2), and Hampel filter (see Section 3.5.3), thereby fostering more effective and robust framework for causal inference in the presence of challenging datasets.

The accurate detection of outliers in a dataset is notably influenced by the underlying data trend. In the context of the dataset with Gaussian noise, a computationally less demanding polynomial interpolation method may suffice to effectively remove the trend. According to the methodology described in Section 3.4, Table 7.1 shows MAD calculations results up to the 9th polynomial order for the given dataset.

TABLE 7.1: Mean absolute deviation (MAD) values for i^{th} polynomial order - dataset with Gaussian noise

	0	1 st	2 nd	3 rd	...	6 th	7 th	8 th	9 th
ϵ_1	0.0536	0.0536	0.0502	0.0478	...	0.0423	0.0412	0.0405	0.0405
ϵ_2	0.0186	0.0186	0.0185	0.0185	...	0.0183	0.0183	0.0183	0.0183
ϵ_3	0.0339	0.0338	0.0338	0.0338	...	0.0338	0.0338	0.0339	0.0340
ϵ_4	0.0355	0.0355	0.0354	0.0354	...	0.0353	0.0353	0.0354	0.0354
ϵ_5	0.0387	0.0387	0.0388	0.0386	...	0.0380	0.0375	0.0366	0.0364

Scrutinizing the results presented in Table 7.1, it becomes evident that elevating the order of the polynomial has a negligible effect on reducing the Mean Absolute Deviation value. As a result, when conducting further investigations on the dataset with Gaussian noise, it is reasonable to adopt order 0 for the polynomial for

each control error ϵ_i (i.e., assuming that the signals are merely shifted by a constant value).

However, in the presence of both Gaussian noise and Cauchy disturbance in the second dataset, a trend becomes more significant, necessitating the adoption of a more sophisticated interpolation technique, such as spline interpolation (see Section 3.4).

It is essential to clarify that the observed trend does not hold for all ϵ_i values in the dataset; rather, it specifically applies to cases, where classical polynomial interpolation leads to a substantial reduction in the Mean Absolute Deviation (MAD) value (see Table 7.2). In particular, this phenomenon is significant in ϵ_1 and ϵ_2 , where the MAD coefficient exhibits a significant decline as the polynomial order increases and reaches a stabilization point at around the 7th order.

TABLE 7.2: Mean absolute deviation (MAD) values for i^{th} polynomial order – dataset with Gaussian noise and Cauchy disturbance

	0	1 st	2 nd	3 rd	...	6 th	7 th	8 th	9 th
ϵ_1	2.5714	2.5947	2.5040	1.6965	...	0.7934	0.6988	0.6895	0.6898
ϵ_2	0.2547	0.2594	0.1813	0.1628	...	0.0512	0.0496	0.0497	0.0495
ϵ_3	0.2367	0.2364	0.2399	0.2399	...	0.2307	0.2294	0.2282	0.2275
ϵ_4	0.0728	0.0726	0.0724	0.0720	...	0.0714	0.0714	0.0712	0.0712
ϵ_5	1.7890	1.8339	1.8414	1.7441	...	1.4454	1.4403	1.3782	1.3392

The utilization of such high-order polynomials, although capable of reducing the MAD for ϵ_1 and ϵ_2 , proves to be impractical from both computational and scientific perspectives. The computational effort associated with the evaluation of high-order polynomials increases significantly, impacting the efficiency of data analysis. Moreover, employing excessively complex polynomial interpolations is not considered a favorable practice in classical scientific methodologies, as it may lead to overfitting and unreliable interpretations of the underlying data.

Figure 7.1 and Figure 7.2 show the outcomes of employing the spline interpolation technique for control errors ϵ_1 and ϵ_2 , respectively, derived from the dataset with Gaussian noise and Cauchy disturbance. Remarkably, a 2nd order polynomial yielded satisfactory results in both instances. This observation is of considerable importance as it parallels the concerns raised earlier for the conventional polynomial interpolation method. The successful implementation of the spline method showcases its efficacy in mitigating the issues associated with the classical approach, obtaining a change from a 7th order polynomial to a 2nd order polynomial.

After preparing the datasets using interpolation methods, they are ready for the outlier detection procedures. However, it is important to acknowledge that each selected method (ESD, IQR, and the Hampel filter) exhibits varying sensitivities to the

Chapter 7. Impact of outliers on the Transfer Entropy method

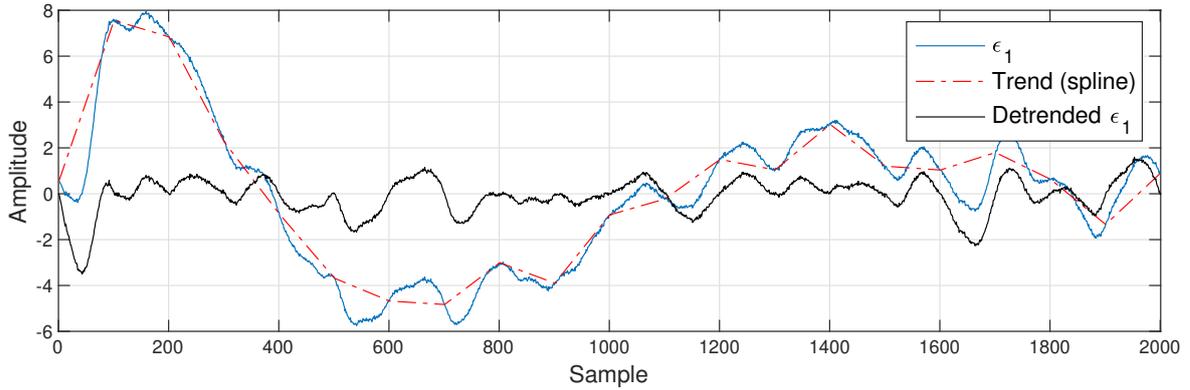


FIGURE 7.1: Results of trend identification and its removal from ϵ_1 using spline interpolation – dataset with Gaussian noise and Cauchy disturbance

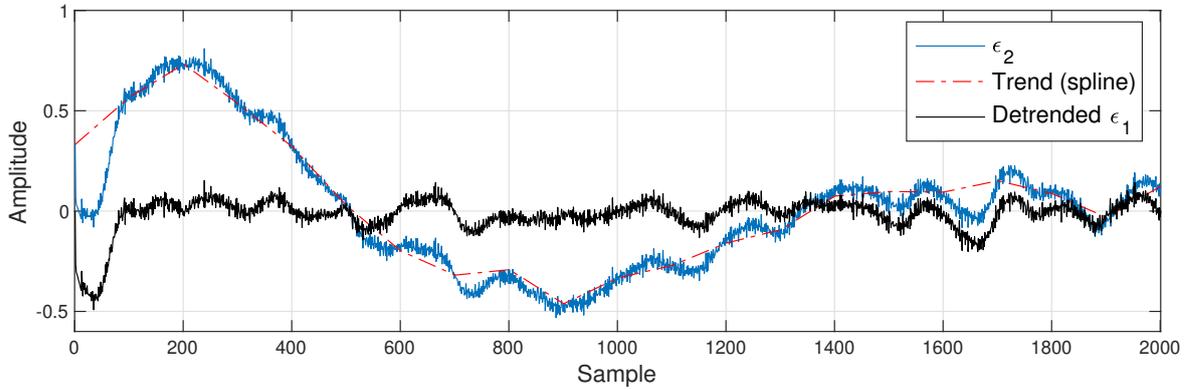


FIGURE 7.2: Results of trend identification and its removal from ϵ_2 using spline interpolation – dataset with Gaussian noise and Cauchy disturbance

type and number of outliers present in the data. Consequently, the lower and upper thresholds determined by these methods differ significantly and impose varying degrees of restrictiveness on the identified outliers.

The results of these outlier identification methods, in the form of *min* and *max* thresholds, for each ϵ_i for both the dataset with Gaussian noise and the dataset with Gaussian noise and Cauchy disturbance are presented in Table 7.3 and Table 7.4, respectively.

In the first analyzed dataset with Gaussian noise, the most effective approach for outlier detection is the Interquartile Range method. For each control error ϵ_i , the minimum and maximum limits are computed based on the IQR, which defines a less restrictive threshold. This characteristic enables the IQR method to identify outliers effectively while preserving the underlying data distribution.

On the other hand, the Hampel filter also calculates threshold values; however, these are considerably more restrictive. If the results from the Hampel filter are

accepted, it may lead to a loss of data dynamics, which is undesirable, particularly from the perspective of the Transfer Entropy method.

The least effective method among the three is the Extreme Studentized Deviate, especially concerning the control errors ϵ_1 and ϵ_2 . Although the computed minimum and maximum values for ϵ_3 , ϵ_4 , and ϵ_5 are higher compared to the IQR method, the absence of results for ϵ_1 and ϵ_2 renders the ESD method unsuitable for further consideration.

Following these observations, it is reasonable to conclude that the IQR method is the most suitable choice for the analysis of the given dataset.

TABLE 7.3: Calculated thresholds values using ESD, IQR and Hampel filter - dataset with Gaussian noise

Control error	ESD		IQR		Hampel	
	<i>min</i>	<i>max</i>	<i>min</i>	<i>max</i>	<i>min</i>	<i>max</i>
ϵ_1	NA	NA	-0.1745	0.1741	-0.0015	0.0021
ϵ_2	NA	NA	-0.0617	0.0619	-0.0093	0.0056
ϵ_3	-0.17805	0.1943	-0.1045	0.1042	-0.0091	0.0057
ϵ_4	-0.21697	0.1900	-0.1175	0.1159	-0.0020	0.0134
ϵ_5	-0.22265	-0.2204	-0.1165	0.1233	-0.0031	0.0058

In both datasets, we obtain identical conclusions. Once again, the application of the Interquartile Range method for the dataset with Gaussian noise and Cauchy disturbance yields objectively satisfactory results, enabling subsequent analysis and the establishment of causal relationships (see Table 7.4).

However, the Hampel filter's threshold values prove excessively restrictive, rendering it unsuitable for utilization. As a consequence, valuable information embedded in the control errors may be lost due to its overly stringent nature.

Likewise, the Extreme Studentized Deviate method displays the lowest effectiveness in this context. Furthermore, it is incapable of determining the minimum and maximum values for control errors ϵ_4 and ϵ_5 . Consequently, it is evident that the ESD method is not the ideal choice for identifying outliers.

TABLE 7.4: Calculated thresholds values using ESD, IQR and Hampel filter - dataset with Gaussian noise and Cauchy disturbance

Control error	ESD		IQR		Hampel	
	<i>min</i>	<i>max</i>	<i>min</i>	<i>max</i>	<i>min</i>	<i>max</i>
ϵ_1	-3.3819	-3.3689	-1.7775	1.8140	-0.1004	-0.0386
ϵ_2	-0.3743	-0.3733	-0.1615	0.15198	-0.0207	-0.0098
ϵ_3	-1.5339	-1.5157	-0.6997	0.7638	-0.0705	0.0067
ϵ_4	NA	NA	-0.2661	0.26508	-0.0038	0.0008
ϵ_5	NA	NA	-4.6939	5.4028	-0.2045	0.2844

7.1. Causality for the dataset with Gaussian noise

The IQR method stands out as the preferred approach for identifying outliers in this dataset during that experiment. It serves as the basis for further investigations into the establishment of causal relationships between control errors. By employing the IQR method, the research endeavors can progress with confidence in the robustness and reliability of the results obtained.

The data processing methodology described above, which includes detrending and identifying outliers using the Interquartile Range method, facilitates the computation of Transfer Entropy coefficients and subsequently enables the construction of causality graphs for both datasets. Emphasizing the significance of these data analysis techniques, it is noteworthy that they allow the calculation of TE coefficients using the classical approach. This advantage can be attributed primarily to the successful attainment of data probability distributions for each control error ϵ_i , which closely approximates the normal distribution. This adherence to normality assumptions enhances the applicability and reliability of the classical TE calculation method, providing valuable insights into causal relationships within the datasets. Moreover, the resulting causality graphs offer a comprehensive visualization of the interdependencies among variables and facilitate the understanding of causal influences governing the simulation system. The results are presented in Section 7.1 and Section 7.2 for each dataset.

7.1 Causality for the dataset with Gaussian noise

This section presents the outcomes of Transfer Entropy computations for a dataset with Gaussian noise. Prior to analysis, the dataset underwent preprocessing, involving trend removal and the application of the Interquartile Range method to identify outliers. The resultant TE coefficients are detailed in Table 7.5.

TABLE 7.5: Calculated Transfer Entropy coefficients using IQR outliers detection method for the dataset with Gaussian noise

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.1207	0.0875	0.0975	0.0800
ϵ_2	0.1162	NA	0.0654	0.0731	0.0676
ϵ_3	0.0662	0.0694	NA	0.0824	0.0755
ϵ_4	0.1594	0.1262	0.0820	NA	0.0711
ϵ_5	0.0668	0.0752	0.0747	0.0997	NA

The causality graph resulting from the TE coefficients presented in Table 7.5 is displayed in Figure 7.3. This graph depicts the inferred causal relationships among the control errors in the given dataset, revealing the directionality and strength of information flow. Additionally, the causality graph is compared to the real graph, which likely represents the true underlying causal connections in the simulation system.

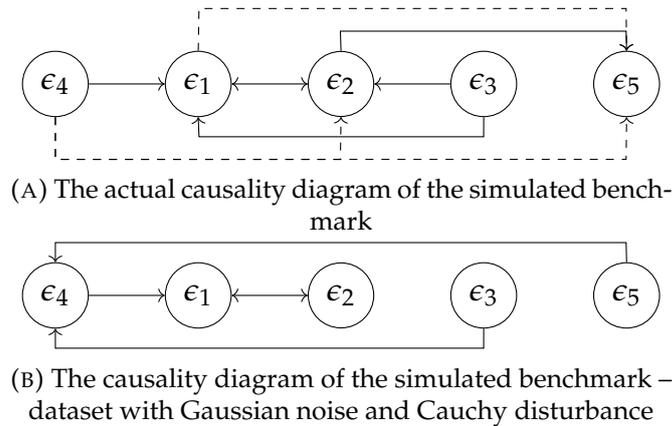


FIGURE 7.3: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance

In the analyzed dataset, the detrending process involved a simple shifting of the samples by a constant value, effectively removing the linear trend component. The IQR method’s determination of limit values significantly influenced the obtained results, successfully identifying outliers and improving the data quality for subsequent analysis.

However, applying the classical Transfer Entropy approach directly to the raw dataset seems to be problematic, making it impossible to present the results in tabular form or as a causality graph. As a consequence, a more refined analysis is required to capture the causal relationships effectively.

The causality graph derived from the TE coefficients exhibited limitations, as it does not accurately represent all the causal connections within the dataset. Although proper causality is evident for relationships between ϵ_1 and ϵ_2 , as well as ϵ_4 and ϵ_1 , other control errors’ causal relationships are incorrect.

It is pertinent to refer to the results presented in Section 6.1.1 and Section 6.2.1, where both parametric and non-parametric statistical models are tested. In this context, the obtained results are less promising, indicating the need for further exploration and refinement of the analysis techniques.

A comparison with the previous results for the dataset with Gaussian noise revealed that while some pairs of control errors yielded meaningful outcomes, the outlier identification process leads to the loss of certain information. Alternatively, fitting the data distribution with less effort, results in more satisfactory and accurate results.

This underscores the importance of choosing appropriate data preprocessing techniques or statistical model selection when conducting causality analyses. Future investigations involve exploring data properties with more sophisticated approaches to improve causality inference accuracy in this case.

7.2 Causality for the dataset with Gaussian noise and Cauchy disturbance

The data processing procedure for the dataset with Gaussian noise and Cauchy disturbance, which involved trend removal through spline interpolation and outlier identification using the IQR method, has a substantial impact not only on the obtained results but also on the underlying nature of the data itself. This observation is particularly relevant to the control errors ϵ_1 and ϵ_2 , as evident in Figure 7.1 and Figure 7.2.

The calculated Transfer Entropy coefficients using the classical approach are presented in Table 7.6, providing quantitative measures of the information flow and causal relationships among the control errors.

TABLE 7.6: Calculated Transfer Entropy coefficients using IQR outliers detection method for the dataset with Gaussian noise and Cauchy disturbance

$T_{row \rightarrow column}$	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	NA	0.2157	0.0415	0.0692	0.0371
ϵ_2	0.2085	NA	0.1842	0.0598	0.0897
ϵ_3	0.0418	0.0805	NA	0.0542	0.0361
ϵ_4	0.0292	0.1282	0.0220	NA	0.0125
ϵ_5	0.0387	0.0705	0.0297	0.0652	NA

The application of spline interpolation for trend removal effectively reduced the influence of trends and fluctuations in the data, resulting in a more stable and consistent dataset. Moreover, the IQR method's outlier identification further enhanced the data quality by identifying and treating potential anomalies and extreme values that could have skewed the analysis.

With the TE coefficients at hand, it becomes possible to construct the causality graph, visualizing the inferred causal connections among the control errors. The causality graph is presented in Figure 7.4b, revealing the directionality and strength of information transfer within the simulation system.

Regrettably, even with the current approach, a complete representation of the existing relationships remains elusive. This is evident in both the inferred causality between ϵ_4 and ϵ_2 , as well as between ϵ_5 and ϵ_4 . When applying the classic Transfer Entropy method to the raw simulation dataset, entirely contradictory results are obtained, leading to their complete rejection (see Section 6.1.2). Nevertheless, a noticeable improvement can be observed when comparing these previous results with the ones currently discussed. The effectiveness of the trend removal and outlier identification algorithms cannot be denied; however, as demonstrated in Section 6.2.2, non-parametric statistical models proved to handle this issue more effectively. Although a reasonable decision is to employ spline interpolation for control errors

ϵ_1 and ϵ_2 , which is reflected in the results, the relationships resulting from the ϵ_4 and ϵ_5 pose challenges.

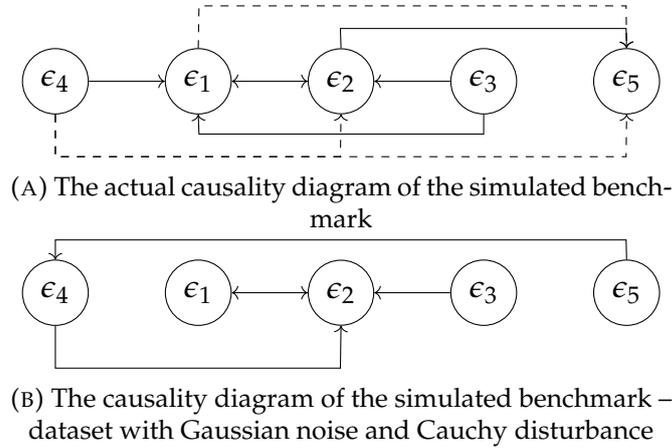


FIGURE 7.4: Comparison of the actual and obtained causality graphs - data impacted by Gaussian noise and Cauchy disturbance

In summary, the research indicates that the choice between parametric and non-parametric statistical models significantly influences the quality of the results. It is essential to emphasize the possibility of combining these methodologies, where trend removal and outlier identification precede fitting the most appropriate probability distribution to the data. However, this approach carries substantial risks.

There is a considerable danger that the impact of these techniques, often subconsciously connected, may lead to significant data distortion, resulting in matching data to methods rather than methods to the data. Such over-interpretation can produce impressive results but remain subjective and applicable only to a specific case, contradictory in other, even similar applications. Being mindful of these limitations, subsequent chapters primarily focus on the nature and properties of the analyzed data, and the selection of the most appropriate method of analysis. The objective is to extract information from the data that might be invisible at first glance, residing within the noise and later within the oscillations. Understanding these intricacies is paramount to uncovering the underlying dynamics and causal relationships within the dataset effectively and objectively.

Chapter 8

Implementation of the Transfer Entropy method for noise signals

Noise signals often play a crucial role in real-world systems, normally introducing uncertainties and complexities within them. Noises can also be indicators of information flow, crucial in root cause analysis. In this chapter, we delve into the causality analysis using the Transfer Entropy approach within a simulation system for the dataset with Gaussian noise and the data set with Gaussian noise and Cauchy disturbance separately. The idea is to use the noise component of the variable as the carrier of the causality information.

Noise signals can be decomposed from the time series through the Ensemble Empirical Mode Decomposition (see Section 3.6.1) or its extension, Multivariate Ensemble Empirical Mode Decomposition (see Section 3.6.2). The study aims to investigate the causality relationships among noises in control errors ϵ_i , and both these methods serve as valuable tools to decompose the signals into Intrinsic Mode Functions (IMFs), enabling a thorough examination of their dynamics and an impact on the Transfer Entropy approach effectiveness.

The EEMD and the MEEMD are data-driven decomposition techniques that offer a flexible and adaptive approach to handling non-linear and non-stationary signals. Both algorithms have their limitations, in particular, the dependence between the length of analyzed data and the number of received IMFs resulting from such a decomposition process.

The main difference between the EEMD and the MEEMD lies in their respective applications to single-dimensional and multi-dimensional data. EEMD is primarily designed for decomposing one-dimensional time series signals into IMFs with different timescales, which collectively capture the signal's non-linear and non-stationary characteristics. MEEMD is specifically tailored for multi-dimensional datasets and can handle multiple interdependent variables or dimensions simultaneously. It decomposes the multi-dimensional data into sets of IMFs for each dimension, thereby capturing the underlying modes of variability and interactions among the variables.

As above, it is decided to apply the MEEMD algorithm to decompose control errors for both datasets and subsequently compute the Transfer Entropy coefficients to construct causality graphs. Investigation of the usefulness of noise in the causality

analysis presents a promising avenue for addressing challenges posed by the data, leading to more robust and accurate causal inference in this case.

Figure 8.1 demonstrates the exemplary effect of decomposition on the control error ϵ_1 from a dataset with Gaussian noise. On the other hand, Figure 8.2 also depicts the decomposition outcome of ϵ_1 , this time from a dataset including both Gaussian noise and Cauchy disturbance.

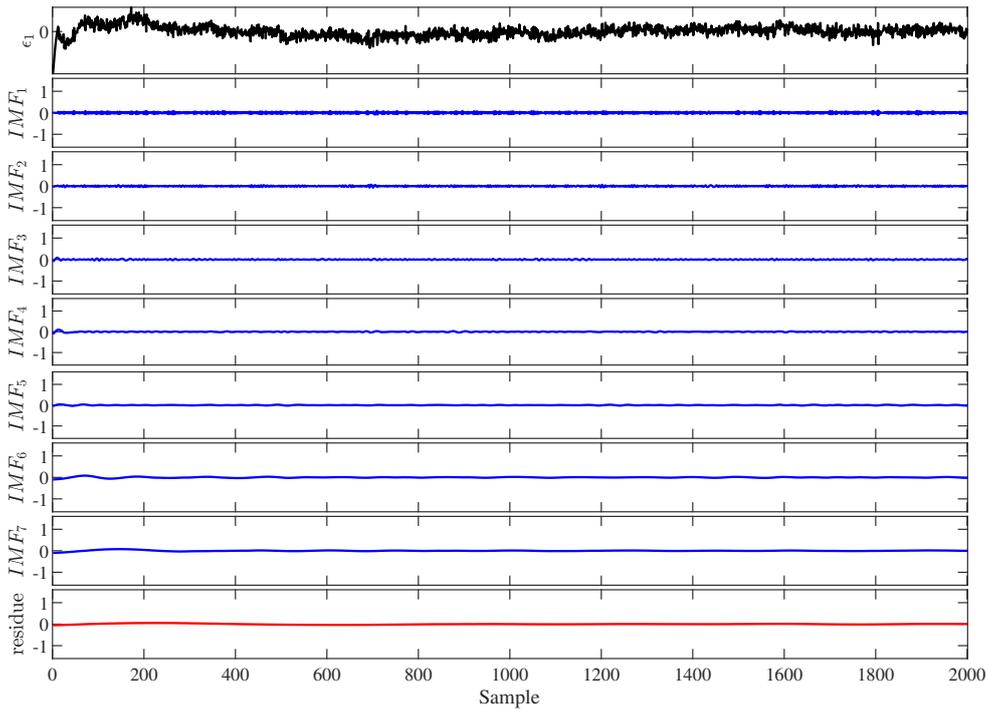


FIGURE 8.1: Results of decomposition process using MEEMD algorithm for control error ϵ_1 – dataset with Gaussian noise

Figure 8.1 and Figure 8.2 present a comprehensive decomposition of the control error ϵ_1 into distinct sections. At the topmost section, the original analyzed signal (control error ϵ_1) is depicted. Below that, seven Intrinsic Mode Functions are shown, representing various intrinsic modes of variability extracted from the data. Again, it is important to note that the number of IMFs obtained through the decomposition process depends on the data's length. In this particular case, the algorithm suggests a decomposition into seven IMFs, each revealing different timescales of variability within the control error signal. Specifically, the IMFs IMF_1 , IMF_2 , and IMF_3 are considered as noise components, while the remaining, i.e., IMF_4 , IMF_5 , IMF_6 , and IMF_7 , are regarded as oscillatory modes. Lastly, at the bottom, the residue is displayed, capturing any remaining fluctuations not captured by the IMFs.

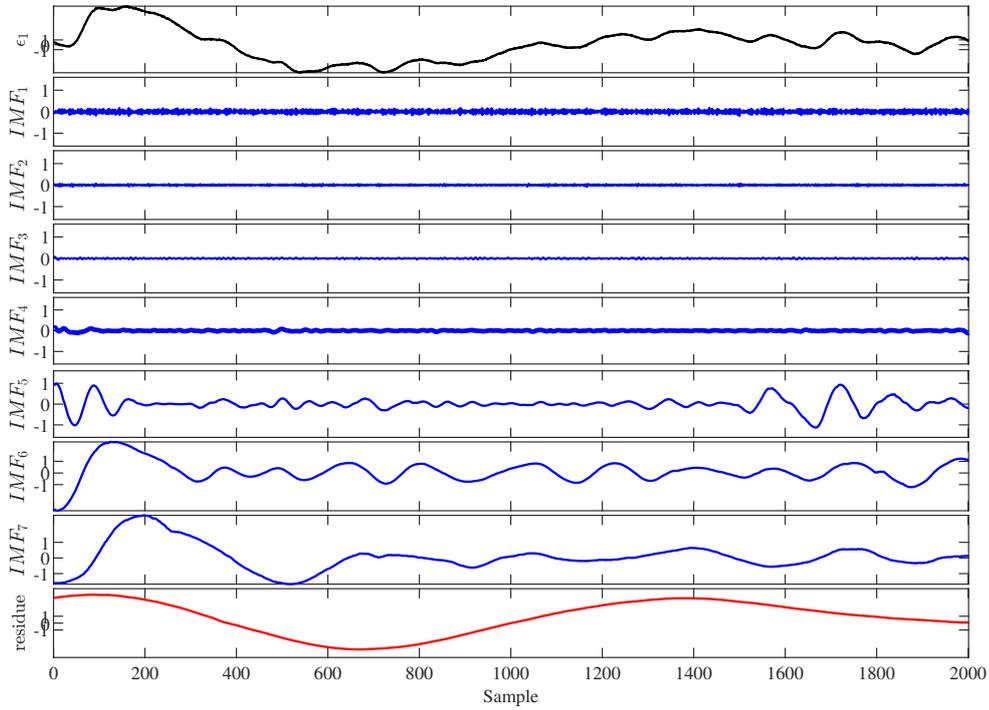


FIGURE 8.2: Results of decomposition process using MEEMD algorithm for ϵ_1 – dataset with Gaussian noise and Cauchy disturbance

The extracted IMFs play a critical role in unraveling the causal relationships between control errors ϵ_i . They allow us to distinguish the noise from significant oscillatory patterns within the data. This functionality is particularly evident in Figure 8.2, where the dataset comprises Gaussian noise and Cauchy disturbance.

By effective isolation of the noise components from the original signal, the IMFs enhance the visibility of underlying oscillations present in the data. The data derived from the decomposition process not only contribute to this Chapter but also have implications for the subsequent Chapter 9.

The subsequent subsections 8.1 and 8.2 present the outcomes of computing Transfer Entropy coefficients for the noise signals extracted from both datasets.

In Section 8.1, we delve into the analysis of TE coefficients for the noise signals derived from the dataset with Gaussian noise. This examination aims at quantification of the directed information flow and causal interactions within the simulation system.

Similarly, Section 8.2 focuses on the TE coefficient calculations for the noise signals obtained from the dataset with Gaussian noise and Cauchy disturbance. This investigation is vital for understanding the causal relationships within a more complex and challenging environment.

At this moment, it is crucial to highlight that the forthcoming tables presenting

the TE coefficients will undergo modifications from their current state. During the decomposition procedure, three distinct noise signals are acquired for each ϵ_i . As a result, the investigation of causality using the Transfer Entropy method is conducted independently for each noise signal, and this process is repeated for every ϵ_i .

This approach ensures that the causality study is comprehensive, allowing for examination of the information flow and directional influences between control error's noises. Such analysis for each noise signal separately allows us to gain a more nuanced understanding of the underlying causal relationships, which may vary depending on the nature and intensity of the noise. These detailed investigations aid in uncovering complex interactions and improving the overall reliability and accuracy of the Transfer Entropy approach.

8.1 Causality for the dataset with Gaussian noise

Table 8.1 exhibits the computed Transfer Entropy coefficients for the noise signals acquired during the Multivariate Ensemble Empirical Mode Decomposition process on the dataset with by Gaussian noise.

TABLE 8.1: Calculated Transfer Entropy coefficients for noise signals – dataset with Gaussian noise

$T_{row \rightarrow column}$		ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	IMF_1	NA	0.5907	0.0529	0.2365	0.0509
	IMF_2	NA	0.1202	0.0846	0.1512	0.0758
	IMF_3	NA	0.1018	0.0956	0.3143	0.0935
ϵ_2	IMF_1	0.2450	NA	0.0545	0.1618	0.0460
	IMF_2	0.4900	NA	0.0735	0.2769	0.0812
	IMF_3	0.3370	NA	0.0938	0.2323	0.1000
ϵ_3	IMF_1	0.0553	0.0499	NA	0.0571	0.0469
	IMF_2	0.0745	0.0656	NA	0.0855	0.0835
	IMF_3	0.0993	0.0997	NA	0.1111	0.0869
ϵ_4	IMF_1	0.2641	0.4543	0.0579	NA	0.0518
	IMF_2	0.1498	0.1035	0.0801	NA	0.0833
	IMF_3	0.2943	0.0989	0.0891	NA	0.0818
ϵ_5	IMF_1	0.0525	0.0510	0.0509	0.0505	NA
	IMF_2	0.0865	0.0720	0.0786	0.0780	NA
	IMF_3	0.1069	0.0913	0.0859	0.0805	NA

The examination of the Transfer Entropy coefficients reveals a notable pattern: despite the successive stages of data decomposition within the presence of Gaussian noise, the results for the noise signals exhibit a degree of consistency. In many

8.1. Causality for the dataset with Gaussian noise

instances, the highest TE values correspond to the same pairs of signals. However, it is essential to note that these relationships between control errors only marginally reflect the actual causal connections within the simulation system.

It is worth noting that the analysis of noise signals reveals specific characteristics of the simulation system, compared with the previous research. Specifically, it is notable during the determination of the relationships involving the control error ϵ_5 . The TE coefficients values computed for ϵ_5 are remarkably similar, suggesting equivalent relationships between this control error and ϵ_1 , ϵ_2 , ϵ_3 , and ϵ_4 . This arises from the inherent characteristic of a closed-loop system, indicating that the ϵ_5 lacks a direct causal relationship with other control errors, as is observable with, for instance, ϵ_1 .

This finding, in conjunction with the previously highlighted issues related to non-singularity and non-unidirectionality in the relationships between specific control errors, is another challenge in determining causality using the Transfer Entropy approach.

In light of these results, Figure 8.3b presents a causality graph to the discussed scenario. This graph encapsulates the observed causal relationships among the control errors within the system, as inferred from the TE coefficients.

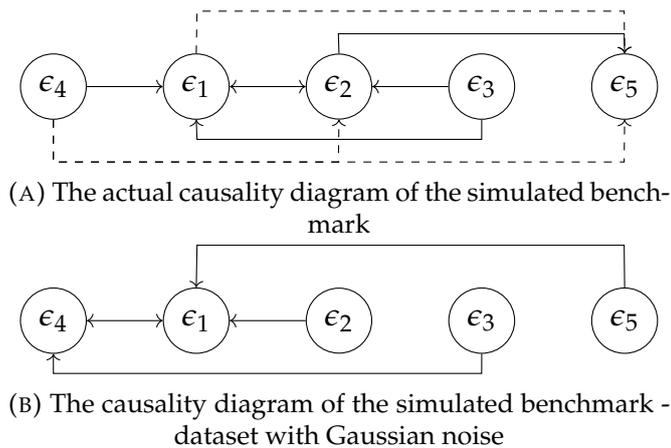


FIGURE 8.3: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise

As previously emphasized, the depicted causal relationships do not accurately reflect the underlying connections through the system. In the majority of instances, the displayed relationships are incorrect, and in some cases, even contradictory. This includes the indirectly inferred relationships, indicated by dashed lines in Figure 8.3a.

Among the various types of signals scrutinized in this research, it becomes evident that the noise within the dataset carries relatively limited information. Consequently, it is not well-suited as a medium for implementing the chosen analytical approach. This observation underscores the importance of selecting appropriate data types and preprocessing techniques when applying causal inference methods,

especially in scenarios where noise components dominate the signal, and their effects may obscure meaningful causal relationships.

8.2 Causality for the dataset with Gaussian noise and Cauchy disturbance

Table 8.2 displays the computed Transfer Entropy coefficients for the noise signals extracted during the Multivariate Ensemble Empirical Mode Decomposition process on the dataset with Gaussian noise and Cauchy disturbance.

TABLE 8.2: Calculated Transfer Entropy coefficients for noise signals – dataset with Gaussian noise and Cauchy disturbance

$T_{row \rightarrow column}$		ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	IMF_1	NA	0.4396	0.0487	0.1903	0.0500
	IMF_2	NA	0.0942	0.0806	0.1085	0.0763
	IMF_3	NA	NaN	0.0961	0.2270	0.1254
ϵ_2	IMF_1	0.1249	NA	0.0448	0.1409	0.0439
	IMF_2	0.3096	NA	0.0841	0.3504	0.0794
	IMF_3	0.2872	NA	0.0890	0.3601	0.1168
ϵ_3	IMF_1	0.0453	0.0407	NA	0.0527	0.0414
	IMF_2	0.0754	0.0808	NA	0.0836	0.0797
	IMF_3	0.0874	NaN	NA	0.0900	NaN
ϵ_4	IMF_1	0.1869	0.4223	0.0487	NA	0.0434
	IMF_2	0.1223	0.0888	0.0835	NA	0.0742
	IMF_3	0.2598	NaN	0.0868	NA	0.1238
ϵ_5	IMF_1	0.0516	0.0466	0.408	0.0462	NA
	IMF_2	0.0846	0.0755	0.0813	0.0760	NA
	IMF_3	0.1234	NaN	NaN	0.1158	NA

Despite the implementation of the Cauchy disturbance, the results shown in Table 8.2 remain highly analogous to those presented in Table 8.1. The sole distinction lies in the depiction of the causal relationship between ϵ_2 and ϵ_4 , which is indicated to be erroneous. This observation implies that, notwithstanding the additional interference factor, the noise signals derived during the dataset decomposition process do not adequately account for this phenomenon. In other words, it appears that the noise signals obtained from the decomposition process do not reflect the distinctive characteristics introduced by the Cauchy disturbance.

Once more, we observe that the Transfer Entropy coefficients pertaining to the control error ϵ_5 exhibit a remarkable degree of proximity, more so than in preceding

8.2. Causality for the dataset with Gaussian noise and Cauchy disturbance

cases. This consistency in values affirms the robustness of the findings derived from the calculations in Section 8.1.

Despite the absence of entirely satisfactory results, this particular observation holds significant value and stands as one of the most pivotal conclusions gleaned from the entire research endeavor. It underscores the persistence of certain causal patterns, even amid the challenges posed by complex noise sources and interference, providing valuable insights for further investigations and analytical refinement.

As in Section 8.1, we obtain inaccurate results. This discrepancy becomes evident upon a careful comparison of the resulting causality graph with the actual representation provided in Figure 8.4.

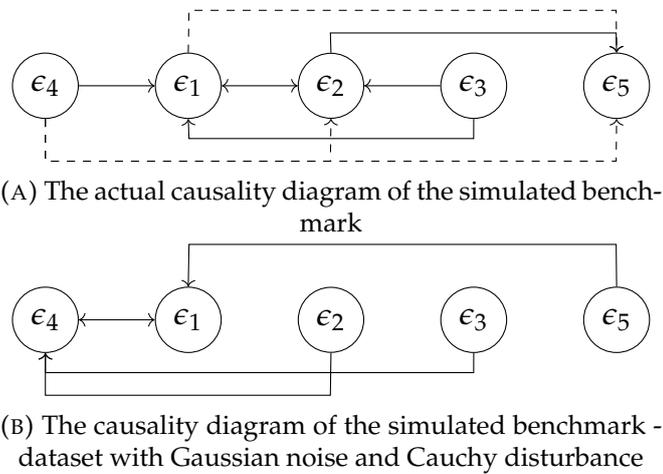


FIGURE 8.4: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance

In reality, all of the inferred causal relationships, with the exception of ϵ_4 and ϵ_1 , are found to be invalid. This observation serves to reinforce the assertion that employing noise as a signal in conjunction with the Transfer Entropy approach significantly diminishes its reliability and efficacy, thus rendering it unsuitable for such applications.

The reproducibility of results across both the dataset with Gaussian noise and the dataset with Gaussian noise and Cauchy disturbance validates the sound functionality of the Transfer Entropy method under the adopted calculation parameters. In both instances, the identified dependencies align; however, their fidelity in representing the actual causal relationships remains notably deficient. This limitation is attributed to the nature of the signals utilized in the computations, specifically, noise signals. The observed limited effectiveness of the method primarily emanates from the intrinsic characteristics of noise itself, which, as evidence, does not convey pertinent information conducive to precise Transfer Entropy analysis. Consequently, it is ascertained that further exploration of this particular case lacks merit, and the inclusion of noise in causal analysis is discouraged.

Chapter 8. Implementation of the Transfer Entropy method for noise signals

As already shown, the absence of satisfactory results does not singularly constitute the sole conclusion drawn from this chapter. The convergence of TE coefficient values for ϵ_5 underscores that the challenges inherent to causal analysis extend beyond the quandary of delineating threshold values to ascertain relationships among control errors, especially when seeking to capture multiple relationships for a given variable. The issue tied to the system's inherent nature, specifically whether it operates as a closed-loop system, emerges as a significant factor. In the context of the subject simulation system, precise knowledge exists regarding which variables lack a direct causal relationship with others. However, this complication is magnified when such knowledge is absent. It is imperative to acknowledge that the introduction of oscillatory signals in the subsequent chapter will not inherently resolve this challenge. Instead, it will serve to evaluate how oscillatory signals, in contrast to previously employed data processing techniques and method variance, impact its accuracy and efficacy. This examination helps shed light on whether oscillatory signals provide a more viable avenue for causal analysis in complex systems.

Chapter 9

Implementation of the Transfer Entropy method for oscillatory signals

Oscillations represent an inherent and ubiquitous component of real signals within complex automation systems. These oscillatory patterns manifest for a multitude of reasons, including but not limited to:

1. **System dynamics**, where the control system's feedback loop generates oscillatory behavior due to its inherent time delays and interactions;
2. **Control tuning**, when a system attempts to regulate itself due to poorly tuned control algorithms or inappropriate control parameters;
3. **Nonlinearities**, such as saturation effects or dead times, can cause oscillations under specific conditions;
4. **Sensor noise**, that can introduce fluctuations in measured signals, potentially leading to oscillations;
5. **Actuators malfunctioning**, like valve stiction;
6. **Resonance frequencies**, when subjected to certain input;
7. **Interference or disturbances**, that are external in most cases;
8. **System instabilities**, resulting from inadequate stability margins or improper design;
9. **Communication delays**, introducing time lags, potentially leading to instability and oscillations;
10. **Controller switching**, that can induce transient oscillations;
11. **Environmental factors**, i.e. temperature fluctuations or other environmental factors that can affect system parameters, leading to oscillations if not adequately controlled.

These are just a few examples, while the specific causes of oscillations can vary significantly depending on the particular system and its operating conditions. Analyzing and mitigating oscillations is a critical aspect of control system engineering to ensure stable and reliable system performance. Despite often being perceived as an undesirable phenomenon, oscillations hold intrinsic value in the context of causal analysis, as they can serve as a rich carrier of transferred information, a crucial element within the framework of the Transfer Entropy method.

The research findings presented in this chapter exhibit a close interrelation with those presented in the preceding Chapter 8. The central focus resides in the data decomposition process, which enables the separation of oscillatory signals from noise components using MEEMD algorithm. This pivotal step lays the foundation for the discussions and investigations in the subsequent Section 9.1 and Section 9.2, offering a structured approach to leveraging oscillatory signals for enhanced causal analysis within the simulation system.

To provide a visual representation of the oscillatory signals extracted through the Multivariate Ensemble Empirical Mode Decomposition process for the pertinent datasets, Figure 9.1 and Figure 9.2 present their waveform characteristics for ϵ_1 , as an example. In this decomposition, each control error ϵ_i is disassembled into seven constituent components. The initial three components are identified as noise components (as elucidated in Chapter 8), while the subsequent components, specifically IMF_4 , IMF_5 , IMF_6 , and IMF_7 , constitute the oscillatory signals that serve as the foundational elements for subsequent research endeavors.

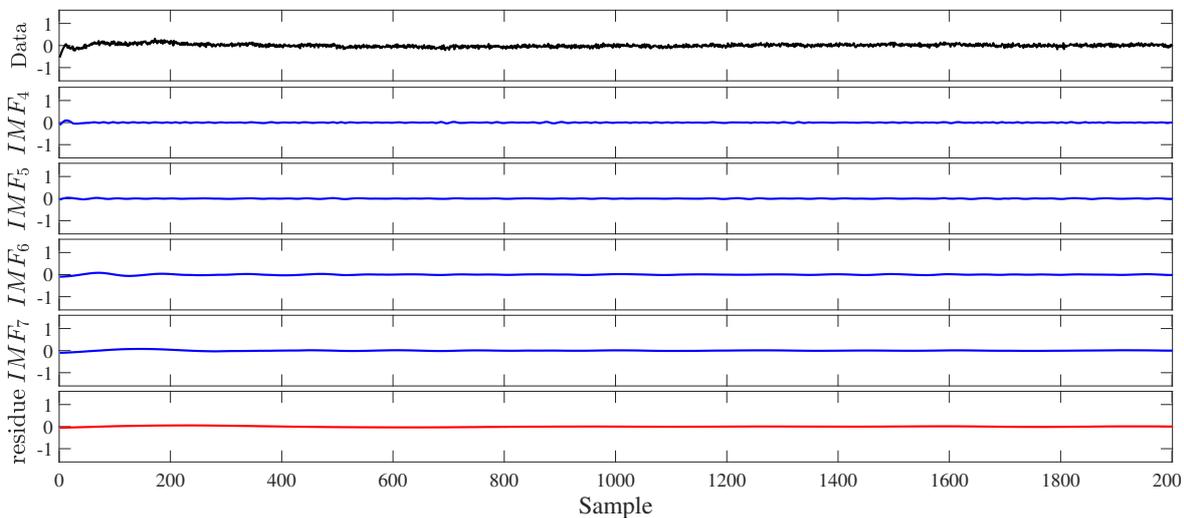


FIGURE 9.1: Oscillatory signals obtained in the process of MEEMD decomposition for control error ϵ_1 – dataset with Gaussian noise

These oscillatory signals, meticulously separated from noise components, form the cornerstone of our investigations, facilitating in-depth analysis and exploration

of their role in elucidating causal relationships within the studied system. Their intricate waveforms encapsulate valuable information that contributes to the broader understanding of system dynamics and causal interactions.

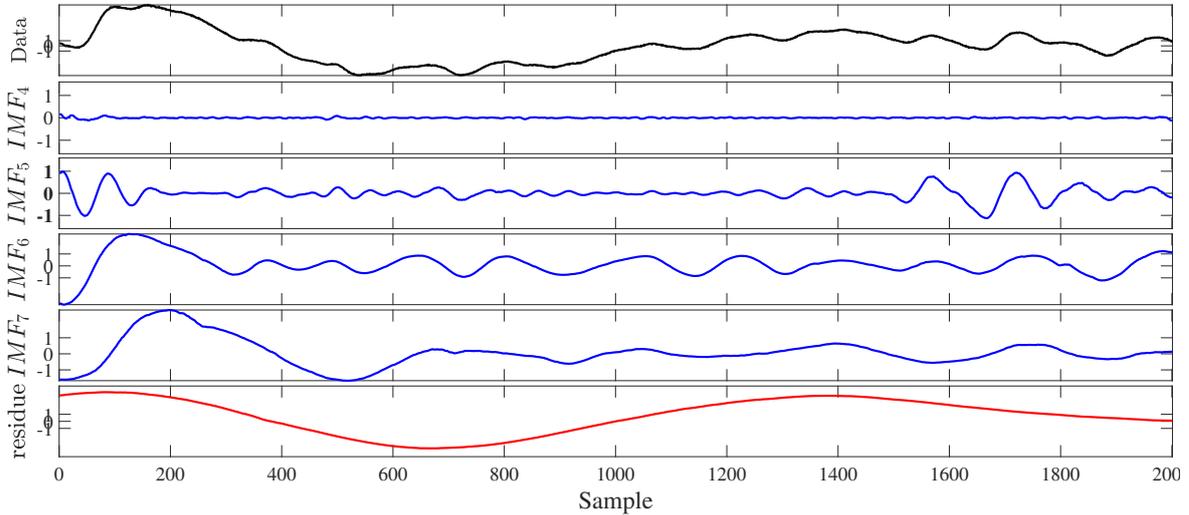


FIGURE 9.2: Oscillatory signals obtained in the process of MEEMD decomposition for ϵ_1 – dataset with Gaussian noise and Cauchy disturbance

It is worth to highlight that in the context of the dataset with Gaussian noise, the properties of the oscillatory signals exhibit significant disparities when compared to those observed in the dataset with Gaussian noise and Cauchy disturbance. This divergence in behavior can be attributed to the limited influence of Gaussian noise on the system. The implementation of Gaussian noise in this particular dataset does not sufficiently perturb the system to induce discernible oscillations.

This observation implies that conducting causality assessments using the Transfer Entropy approach on oscillatory signals within the dataset with Gaussian noise may yield not only inconclusive but potentially misleading results. In essence, the absence of clear oscillatory patterns diminishes the likelihood of these signals conveying information relevant to the Transfer Entropy method.

Conversely, the situation is markedly more promising in the case of the dataset with Gaussian noise and Cauchy disturbance. Here, oscillations are very distinct, and consequently, the likelihood of these oscillatory signals containing pertinent information for the Transfer Entropy method is notably higher. This suggests that the utilization of oscillatory signals in causal analyses within the context of the dataset with Gaussian noise and Cauchy disturbance holds greater potential for yielding meaningful insights into the underlying causal relationships within the simulation system.

Similarly to the analysis that uses noise signals, Section 9.1 and Section 9.2 are dedicated to showcasing the outcomes of Transfer Entropy coefficient calculations,

specifically pertaining to the dataset with Gaussian noise and the dataset with Gaussian noise and Cauchy disturbance, respectively. In these sections, an investigation into relationships is undertaken for each of the individual IMF_i components. The primary objective here is to ascertain the consistency and reproducibility of the results across the various oscillatory components.

By the systematical analysis of the Transfer Entropy coefficients for each IMF_i , we aim to discern any recurring patterns or causal relationships that may exist within these oscillatory signals. This comprehensive approach allows us to evaluate the robustness of the findings and gain deeper insights into the causal dynamics embedded within the oscillatory components of both datasets.

9.1 Causality for the dataset with Gaussian noise

Table 9.1 displays the Transfer Entropy coefficient values pertaining to the oscillatory signals, derived through the decomposition process of the dataset with Gaussian noise, for each respective ϵ_i control error.

As anticipated, the presented outcomes do not offer substantial insights. In most instances, regardless of the depth of decomposition, obtaining Transfer Entropy coefficient values was unfeasible. This behavior is inherently linked to the previously highlighted issue of insufficient system excitation, resulting in a lack of observable oscillations. Consequently, constructing a causality graph based on these findings proves unattainable. One may contemplate the significance of these *NaNs* encountered in the computational process. Nevertheless, this matter delves into the realm of data processing, warranting dedicated investigation in forthcoming research.

However, it is imperative to recognize that the challenges encountered in this analysis do not inherently invalidate the pursuit of causal analysis utilizing oscillatory signals. In this particular instance, the decomposition process succeeded primarily in isolating oscillations from noise, yet these oscillations, in contrast to noise, appear to carry even less, if any, meaningful information.

Upon closer examination, a significant observation emerges from the analysis. Despite the application of the TE method across successive IMF_i components, coefficient values are not attainable. This raises a pertinent question: what if we were dealing with a single waveform of an oscillatory signal, and the TE method still failed to yield coefficient values? To elaborate further, it is highly plausible that even when we possess individual oscillatory signals for various control errors, the calculation of coefficients for certain pairs may remain elusive. This should not necessarily imply an absence of causal relationships between the particular control loop and others within the intricate context of a multi-loop system. This phenomenon becomes evident through the calculations conducted on the dataset with Gaussian noise and Cauchy disturbance, as observed in Section 9.2.

9.2. Causality for the dataset with Gaussian noise and Cauchy disturbance

TABLE 9.1: Calculated Transfer Entropy coefficients for oscillatory signals – dataset with Gaussian noise

$T_{row \rightarrow column}$		ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	IMF_4	NA	NaN	NaN	0.3084	0.1397
	IMF_5	NA	NaN	NaN	0.4890	0.1980
	IMF_6	NA	NaN	NaN	0.2636	0.2462
	IMF_7	NA	NaN	NaN	0.1010	0.0844
ϵ_2	IMF_4	NaN	NA	NaN	0.1407	0.1246
	IMF_5	NaN	NA	NaN	NaN	NaN
	IMF_6	0.2927	NA	NaN	0.2076	NaN
	IMF_7	NaN	NA	NaN	0.1236	0.1606
ϵ_3	IMF_4	NaN	NaN	NA	NaN	0.1395
	IMF_5	NaN	NaN	NA	NaN	NaN
	IMF_6	0.3068	NaN	NA	0.2347	NaN
	IMF_7	0.1545	NaN	NA	0.1079	0.2093
ϵ_4	IMF_4	NaN	NaN	NaN	NA	0.1395
	IMF_5	0.4765	NaN	NaN	NA	0.1478
	IMF_6	0.4100	NaN	NaN	NA	0.1560
	IMF_7	0.1138	NaN	NaN	NA	0.1031
ϵ_5	IMF_4	NaN	Nan	NaN	0.1051	NA
	IMF_5	0.2642	NaN	NaN	0.1956	NA
	IMF_6	0.2632	NaN	NaN	0.2509	NA
	IMF_7	0.1991	NaN	NaN	0.1327	NA

9.2 Causality for the dataset with Gaussian noise and Cauchy disturbance

Table 9.2 presents the Transfer Entropy coefficient values associated with the oscillatory signals from the dataset with Gaussian noise and Cauchy disturbance.

The analysis of Transfer Entropy coefficient values in this context reveals a notable degree of consistency. Irrespective of the level of decomposition denoted by IMF_i , the highest coefficient values consistently correspond to the same pairs of control errors. An exception arises with IMF_4 , where distinct dependencies emerge, or coefficient values calculated at this decomposition level cannot be expressed numerically. This observation engenders a crucial conclusion.

The Multivariate Ensemble Empirical Mode Decomposition method has effectively disentangled noise from oscillations within the original dataset, as evidenced in Chapter 8. While applying the Transfer Entropy method to noise signals yielded

TABLE 9.2: Calculated Transfer Entropy coefficients for oscillatory signals – dataset with Gaussian noise and Cauchy disturbance

$T_{row \rightarrow column}$		ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5
ϵ_1	IMF_4	NA	NaN	NaN	0.1367	0.1685
	IMF_5	NA	0.3388	NaN	0.1283	0.2596
	IMF_6	NA	0.3683	0.3310	0.1630	0.1528
	IMF_7	NA	0.2254	0.1785	0.0957	0.1464
ϵ_2	IMF_4	NaN	NA	NaN	0.1191	0.1305
	IMF_5	0.6457	NA	0.3246	0.1305	0.2504
	IMF_6	0.3767	NA	0.3017	0.1899	0.1469
	IMF_7	0.1590	NA	0.0675	0.1118	0.1224
ϵ_3	IMF_4	0.3356	NaN	NA	0.0966	0.1911
	IMF_5	0.3804	0.3761	NA	0.1905	0.2611
	IMF_6	0.1895	0.0611	NA	0.0817	0.1754
	IMF_7	0.1430	0.1052	NA	0.1246	0.1026
ϵ_4	IMF_4	0.1887	NaN	0.1126	NA	0.1625
	IMF_5	0.1703	0.1524	0.1300	NA	0.1488
	IMF_6	0.1977	0.1378	0.1514	NA	0.1312
	IMF_7	0.1291	0.1140	0.1255	NA	0.1247
ϵ_5	IMF_4	0.1582	NaN	0.1731	0.1192	NA
	IMF_5	0.2353	0.2550	0.1804	0.1669	NA
	IMF_6	0.1730	0.2011	0.1269	0.01529	NA
	IMF_7	0.1302	0.1526	0.0875	0.1430	NA

moderate results, its application to oscillatory signals has proven highly satisfactory. However, the pivotal consideration lies in the level of data decomposition, which directly influences the number of oscillatory signals generated.

Assuming four levels of decomposition for the examined control errors, erroneous results would manifest for IMF_4 , rendering the selection of oscillations for TE analysis redundant. Hence, it is imperative to determine an optimal and objectively effective number of Intrinsic Mode Functions for each control error's decomposition. This ensures not only the acquisition of valuable analytical results but also avoids superfluous computational overhead. Furthermore, this analysis facilitates the comparison of results across multiple levels of decomposition, confirming the consistency of the derived conclusions.

Figure 9.3 showcases the causality graph for this scenario, offering a comparative perspective against the current representation.

As anticipated, the causality graph derived from the dataset with Gaussian noise and Cauchy disturbance, based on oscillatory signals, closely reflects the actual system's causal relationships. It effectively captures the genuine connections between

9.2. Causality for the dataset with Gaussian noise and Cauchy disturbance

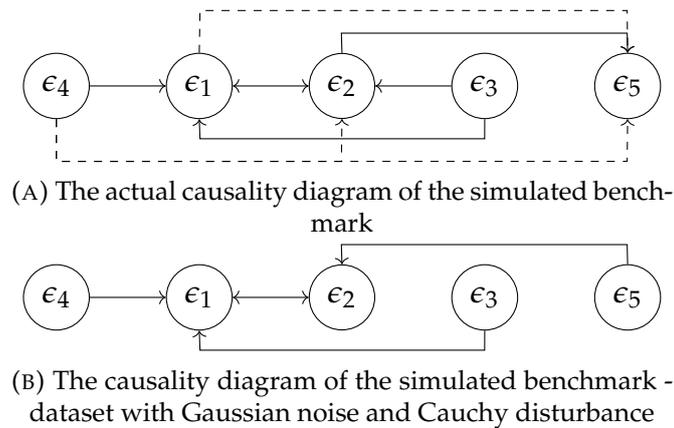


FIGURE 9.3: Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance

the specified control errors. These findings underscore the caveats highlighted in earlier investigations.

Firstly, the limitation imposed by the selected criterion for determining relationships (i.e., the maximum TE value within a given set of pairs) results in the inability to ascertain multiple connections for a single error. Secondly, the closed-loop nature of the analyzed simulation system introduces complexities. It is exemplified by the calculated relationship between ϵ_5 and ϵ_3 , which, while correct in existence, exhibits an inaccurate direction.

The comprehensive research process has facilitated an in-depth exploration of both the dataset and the Transfer Entropy method as a tool for causality analysis. Numerous facets, encompassing not only the inherent system dynamics but also data processing factors influencing the method's efficacy, have been meticulously examined. Chapter 10 presents an exhaustive account of the insights garnered thus far. This knowledge forms the basis for the development of an objective analytical procedure designed to execute effective causal analyses within complex multi-loop control systems.

Chapter 10

Research overview and industrial validation

10.1 Summary

During the investigation, we have examined a simulation system subject to testing in two distinct variants. In the first scenario, we scrutinize a dataset affected by the Gaussian noise. The analysis focuses on control errors, which exhibit a relatively stationary character, and their distribution can be considered close to the normal one. In the second case, we implement fat-tailed Cauchy disturbance simulating industrial-like impacts, markedly inducing serious oscillatory behavior to the control errors. The probability distributions of this dataset deviate from the Gaussian one. We consistently apply the same data processing techniques to both variants, combined with various modifications to the Transfer Entropy method. The objective is to underscore the pivotal role played by the characteristics and dynamics of the data selected for analysis. This approach not only enables the application of objective data analysis methodologies but also engenders significant insights into the employed analytical approach.

Evidently, the Transfer Entropy method does not have to be confined solely to humanities research, as exemplified here, but extends its utility to the realm of technical disciplines, specifically to control engineering. Nonetheless, it exhibits noteworthy limitations and requires prerequisites that must be satisfied for its reliability and resilience in accommodating the characteristics of the analyzed problem. It is pertinent to highlight the judicious choice of data for the analysis, in this instance, control errors, as previously substantiated in Section 3. These limitations can be categorized into two distinct domains: those arising from the inherent nature of the system itself and, consequently, the dataset's characteristics, and those stemming from the chosen Transfer Entropy methodology.

The initial category of constraints revolves around the inherent inability to conduct an objective assessment of the system using only raw data. The intricate informational content embedded within the analyzed datasets hampers the attainment of precise outcomes and, consequently, the accurate delineation of relationships among control errors. Presuming the suitability of the classic Transfer Entropy approach,

founded on probability calculations contingent upon a normal distribution, is fundamentally flawed. This is primarily attributed to the fact that, in the majority of cases, we confront nonlinear systems and data characterized by a non-Gaussian distribution. Hence, an exhaustive scrutiny of the datasets is imperative, accompanied by their meticulous preprocessing.

This ushers in the second constraint, concerning data processing. Raw datasets often fail to yield the desired outcomes, when subjected to the conventional TE methodology. Consequently, further steps involving data manipulation or the segregation of signal components are contemplated, contingent upon the specific nature of the data. Although theoretically promising, it is crucial to exercise caution, when selecting data processing algorithms. The main consideration is to prevent data from being forcibly aligned with the algorithm, engendering an artificial enhancement in TE efficiency that is context-specific. However, it is imperative to refrain from the wholesale exclusion of these techniques since, in real control systems, they may significantly influence the quality of TE analysis.

Additionally, once we deal with a relatively unexcited system, the efficacy of data processing notably diminishes, often resulting in misleading outcomes and consequent dependencies (the well-known identification requirement for enough excitation of the variables). For such data, solutions should be sought in the versatility of the TE method, as manifested in the successful alignment of probability distribution types with control errors (see Section 6). In systems exhibiting pronounced oscillatory characteristics, the manipulation of oscillatory signals derived from the decomposition process yields substantial benefits. Nonetheless, the outcomes are not without imperfections, leading to an additional inference.

Causality analysis in the context of close-loop control systems presents inherent challenges. As shown in this research, it is feasible to capture relationships within a control loop, even when the output signal of that loop lacks a direct connection with another loop. However, the issue arises in discerning the correct direction of this relationship. When conducting research solely based on data without prior knowledge of the system, it can potentially yield a false representation of the system's structure, as gleaned from calculations using the Transfer Entropy method. It is conjectured that in closed-loop systems, this phenomenon is less likely to occur, as each signal typically maintains a close relationship with others within the system.

One of the most significant drawbacks of the Transfer Entropy, among those associated directly with the method itself, is the absence of an objectively determined threshold for establishing relationships between control errors. This point has been reiterated throughout this research. In numerous instances, the TE values obtained are very similar, rendering the strict criterion of selecting the highest coefficient value as the most objective but not entirely sufficient. Consequently, the only recourse is to assume a solitary relationship between the pairs of examined control errors. This approach precludes the determination of dependencies involving one loop and two or more others, a common occurrence in multi-loop control systems. Consequently, the system's depicted structure is simplified to the utmost degree.

10.1. Summary

However, this simplicity may overlook crucial relationships that hold significant implications for the overall process.

Another noteworthy conclusion drawn from this research is the computational efficiency of the Transfer Entropy method when applied to extensive long datasets. To manage longer time series, one might consider dividing them into smaller subsets. However, this approach carries the risk of potential losses of necessary information that is relevant to the TE method. Hence, caution should be exercised during such practices.

The aforementioned doubts, deliberations, and conclusions collectively contribute to the formulation of the original procedure for conducting causality analysis of the multi-loop control systems using the Transfer Entropy method. The proposed procedure is visually represented in Figure 10.1, providing a structured guideline for investigations in the causality analysis of complex control systems. The proposed **causality assessment procedure**, which is shown graphically and described narratively below constitutes the synthetic result of the performed research.

The initial phase of the **causality assessment procedure** focuses on the acquisition of imperative data for subsequent analysis, specifically focusing on control error variables derived from the considered system. Control errors serve as crucial indicators of system performance and behavior, necessitating a meticulous examination.

Step 1 Data Acquisition and Time-Domain Analysis

Data Acquisition: The primary step entails the acquisition of control error data originating from the target control system. These variables encapsulate vital information about the control system's performance.

Time-Domain Analysis: Subsequently, an assessment of the time characteristics of these time series is conducted. The aim is to ascertain the presence or absence of oscillations within the system dynamics, as they often mean poor tuning, nonlinearities, or instability.

Step 2 Analysis Pathways Depending on the Decomposed Oscillation

Case 1 Absence of Oscillations

In scenarios, where the oscillations are not evident in the analyzed dataset, a specific set of analytical steps is initiated:

Trend Removal: If any underlying trends exist in the control error data, these trends are removed. This action ensures that the focus remains on the intrinsic characteristics of the trend stationary data.

Histogram Analysis: Following trend removal, the control error data is subjected to histogram analysis. This facilitates the identification of potential outliers and offers further insight into the data.

PDF Fitting: If the data distribution diverges from normality, a suitable parametric or non-parametric probability distribution is fitted to the data. This distribution forms the foundation for subsequent calculations.

Transfer Entropy Causality Analysis: Utilizing the chosen probability distribution, Transfer Entropy analysis is performed. Transfer Entropy quantifies the causality relationships within the data, ultimately leading to the construction of the causality graph.

Case 2 Presence of Oscillations

In scenarios where the system exhibits clear oscillations, the following steps are integrated into the procedure:

Decomposition Assessment: The determination of the decomposition level becomes pivotal, especially when employing the Multivariate Empirical Mode Decomposition approach. If the decomposition level exceeds 7, the dataset necessitates subdivision into smaller, manageable subsets.

MEEMD Decomposition: Employing the MEEMD methodology, the dataset undergoes a systematic decomposition process. This process disentangles oscillatory components from the original data, effectively isolating the embedded noise.

Transfer Entropy Causality Analysis: Post-decomposition, the causality analysis is conducted using Transfer Entropy. This analysis reveals causal relationships within the decomposed oscillatory components.

Step 3 Causality Graph Presentation

The outcome of the analysis is presented in the form of the causality graph, showing fundamental causality relationships within the system.

Concluding, this procedure offers an original, structured, and scientifically-grounded, but practical approach for the comprehensive analysis of causality relationships in the multi-loop control systems, tailored to the presence or absence of oscillations (poor-tuned or well-tuned system). The methodology encompasses data preprocessing, stochastic analysis, and advanced techniques to unveil causal relationships.

The proposed original procedure is derived using simulation and theoretical analysis. Its practical potential is shown using real industrial data. The consecutive Section 10.2 provides a comprehensive causality analysis example, which is executed according to the proposed procedure shown in Figure 10.1. It includes the industrial validation of the proposed **causality assessment procedure** run on real data from the ammonia synthesis system of the Grupa Azoty, Zakłady Azotowe "Puławy" SA.

10.1. Summary

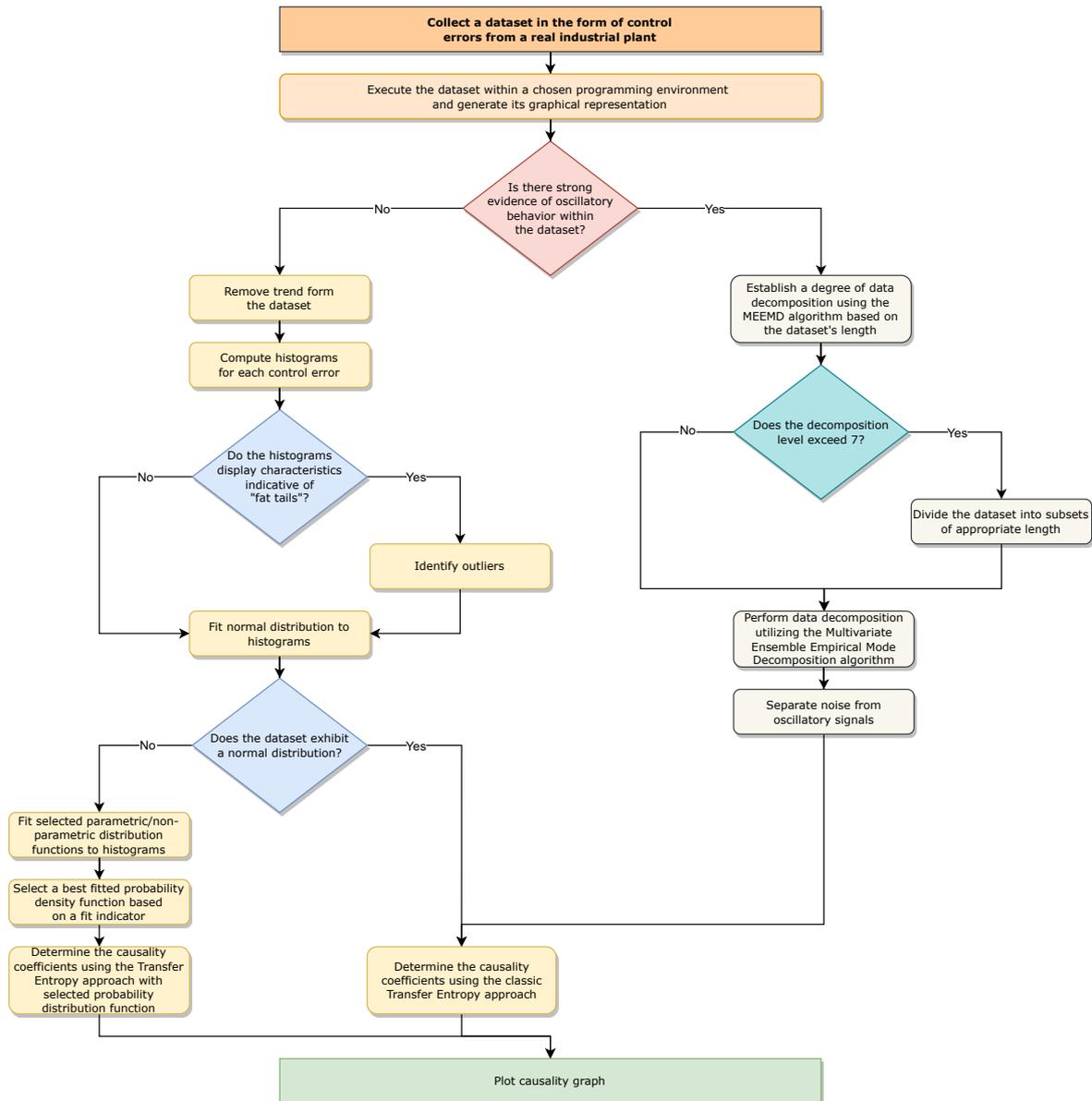


FIGURE 10.1: Diagram of the proposed analytical procedure for causality analysis using the Transfer Entropy approach

10.2 Validation with the ammonia synthesis installation dataset

The subsequent section contains the results of the causal analysis conducted on the dataset encompassing control errors within the ammonia synthesis installation. According to the **causality assessment procedure**, the causality graphs are derived as a consequence of systematically executing the procedure in the specified sequence. The process unfolds in the following order: Step 1, followed by Step 2 (specifically, Case 1), and ultimately, Step 3. This sequential approach ensures that each step builds upon the results and insights obtained from the previous one, facilitating a comprehensive assessment of causality relationships within the dataset. This assertion is corroborated by the insights derived from the analyses detailed in Section 4.2.

It is important to emphasize that the outcomes originally derived from the simulation system, as illustrated by the table showcasing Transfer Entropy coefficients and the accompanying causality graph, do not hold direct applicability within the present context. This divergence is chiefly attributed to the substantial volume of datasets, as well as the multifarious control errors for which causal relationships are scrutinized, as elaborated upon in Section 4.2. In strict adherence to the **causality assessment procedure** delineated in Figure 10.1, there are presented a total of 14 distinct causality graphs, each corresponding to one of the 14 consecutive months encompassing the operational trajectory of the ammonia synthesis installation.

In contrast to the comprehensive causality graph depicted in Figure 4.23, each of the corresponding graphs presented in Figure 10.2, Figure 10.3, Figure 10.4, Figure 10.5, Figure 10.6, Figure 10.7, Figure 10.8, Figure 10.9, Figure 10.10, Figure 10.11, Figure 10.12, Figure 10.13, Figure 10.14, and Figure 10.15, includes additional nodes representing ambient (outdoor) parameters, specifically air temperature (T_α), air density (ρ_α), humidity level (H), and atmospheric pressure (P_e). The purpose of this investigation is to check the potential influence of external factors on the observed control errors, thereby establishing the interrelationships between these environmental variables and the control errors in question. This approach allows for a more holistic understanding of how ambient conditions might contribute to variations in relationships between control errors.

It must be noted that the data used for the causality analysis are reviewed and selected very carefully. This step, though rarely included in the research reports, is inevitable for the analysis. Poorly chosen or incomparable time series lead to wrong conclusions. Data selections must lead to a reliable and comparable time series. They must refer to the similar operating conditions of the installation. Considered analysis uses data from 14 months of operation (07/2020–08/2021), which are sampled with 1-minute time intervals. Much attention is paid to finding installation operation periods characterized by comparable loads and similar operating conditions. Plant load varies $\pm 1.5\%$ in the considered data. The load demand within

10.2. Validation with the ammonia synthesis installation dataset

the data is not intentionally affected by existing variations related to normal process behavior. Moreover, the data do not include periods with known hardware issues.

The analysis of causality graphs derived from the 14-month operational data of the ammonia synthesis installation reveals a remarkably high degree of similarity among them. Specifically, two distinct groups of identical relationships between control errors have been identified. The first group encompasses months 1, 4, 5, 6, and 12, while the second group includes months 2, 3, 7, 8, 9, 10, and 11. The sole distinction between these obtained graphs lies in the alternating causal relationship, with the first group featuring a connection between $\epsilon_{F_{12}}$ and ϵ_{L_1} , and the second group showing a connection between $\epsilon_{F_{12}}$ and ϵ_{L_2} . It is imperative to underscore that all established relationships between control errors are consistent in both cases.

A significant observation arising from the comparative analysis between the existing causality graph presented in Figure 4.23 and the derived causality graphs refers to the nature of the relationship between variables ϵ_{F_7} and ϵ_{T_3} with respect to the $\epsilon_{F_{10}}$. For all 14 periods, employing the Transfer Entropy method for calculations yielded the highest Transfer Entropy coefficients for these errors. This implies that within the resultant causality graphs, the associations involving the control errors ϵ_{F_7} and $\epsilon_{F_{10}}$, as well as ϵ_{T_3} and $\epsilon_{F_{10}}$, are considered direct in nature.

In contrast to the actual causality graph depicted in Figure 4.23, the subsequent graphs also incorporate causal relationships involving ambient temperature (T_α), atmospheric pressure (P_e), humidity level (H), and air density (ρ_α). As expected, these variables exhibit close and reproducible causal connections. Notably, in none of the cases does a relationship with any of the control errors emerge, irrespective of whether the focus is on errors originating from control loops related to flow, level, pressure, or temperature.

The ultimate conclusion of the above analysis confirms the proposed analytical methodology. Despite a long period of the analysis, which covers 14 months of the plant operation the analysis delivers comparable and repeatable results. Moreover, the obtained causality graphs are consistent with the real relations. This proves the method's reliability.

A limitation encountered in the application of the Transfer Entropy method observed during the analysis of the simulation system, and likewise encountered in the analysis of industrial data, pertains to its incapacity to discern multiple relationships involving a single control error. In many instances, we encountered actual dependencies where one control error exhibited associations with several others, as exemplified by ϵ_{F_3} or ϵ_{F_6} . This constraint arises from the underlying assumption that the relationship between control errors is established based on the highest value of the Transfer Entropy coefficient. A potential remedy for this limitation is elaborated upon in the Conclusion (Chapter 11).

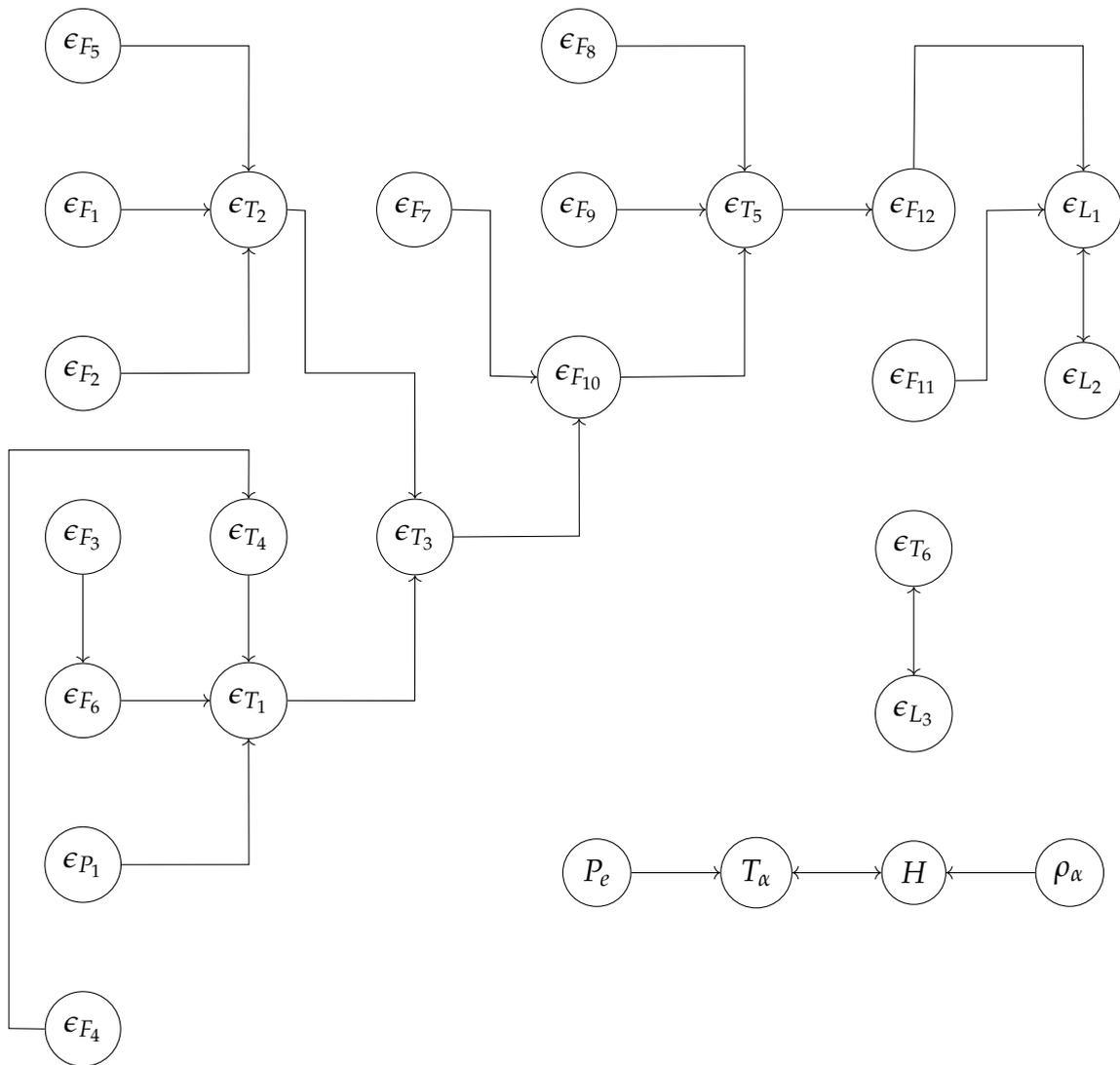


FIGURE 10.2: The causality diagram of the ammonia synthesis installation – 1st month dataset

10.2. Validation with the ammonia synthesis installation dataset

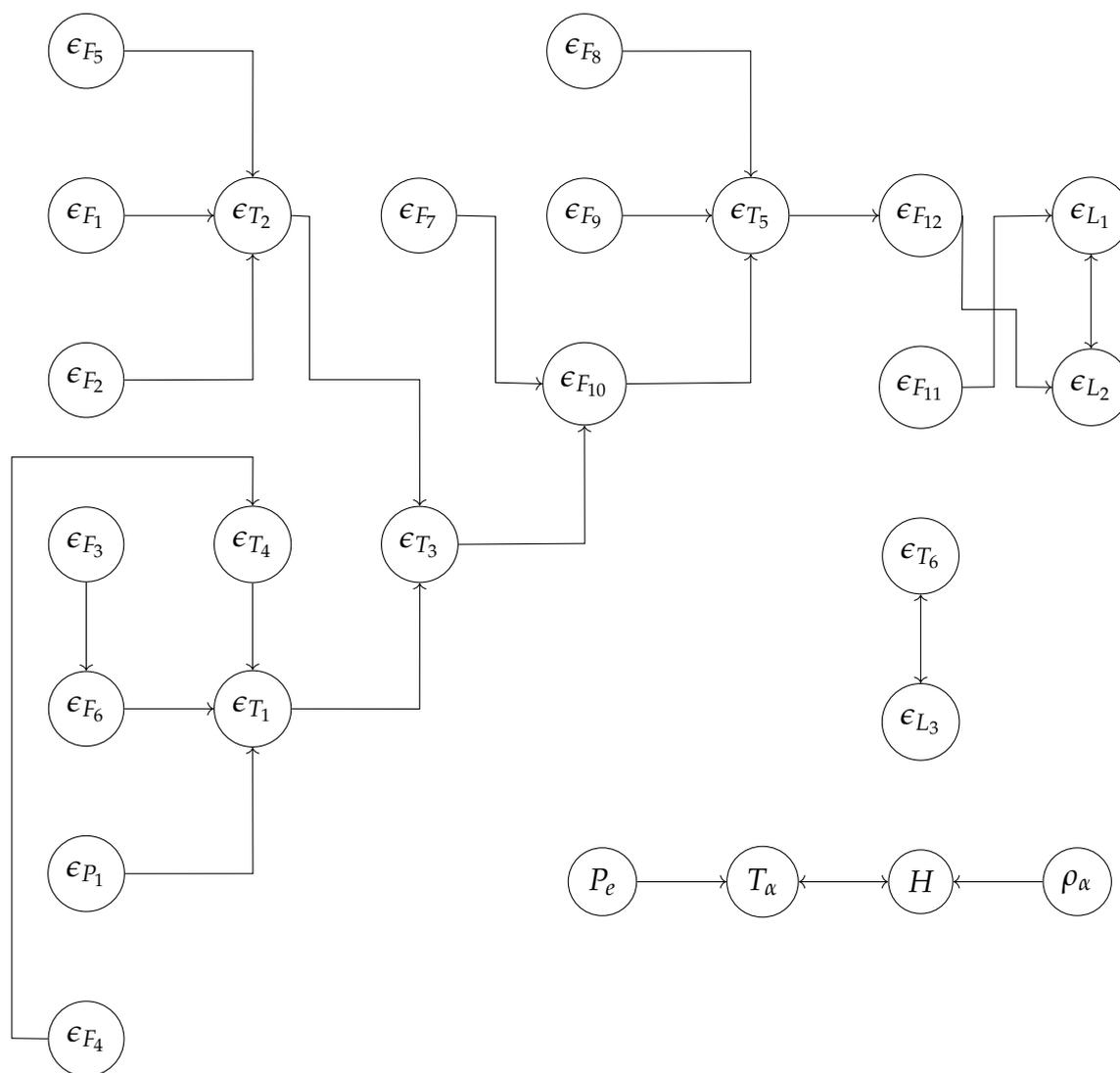


FIGURE 10.3: The causality diagram of the ammonia synthesis installation – 2nd month dataset

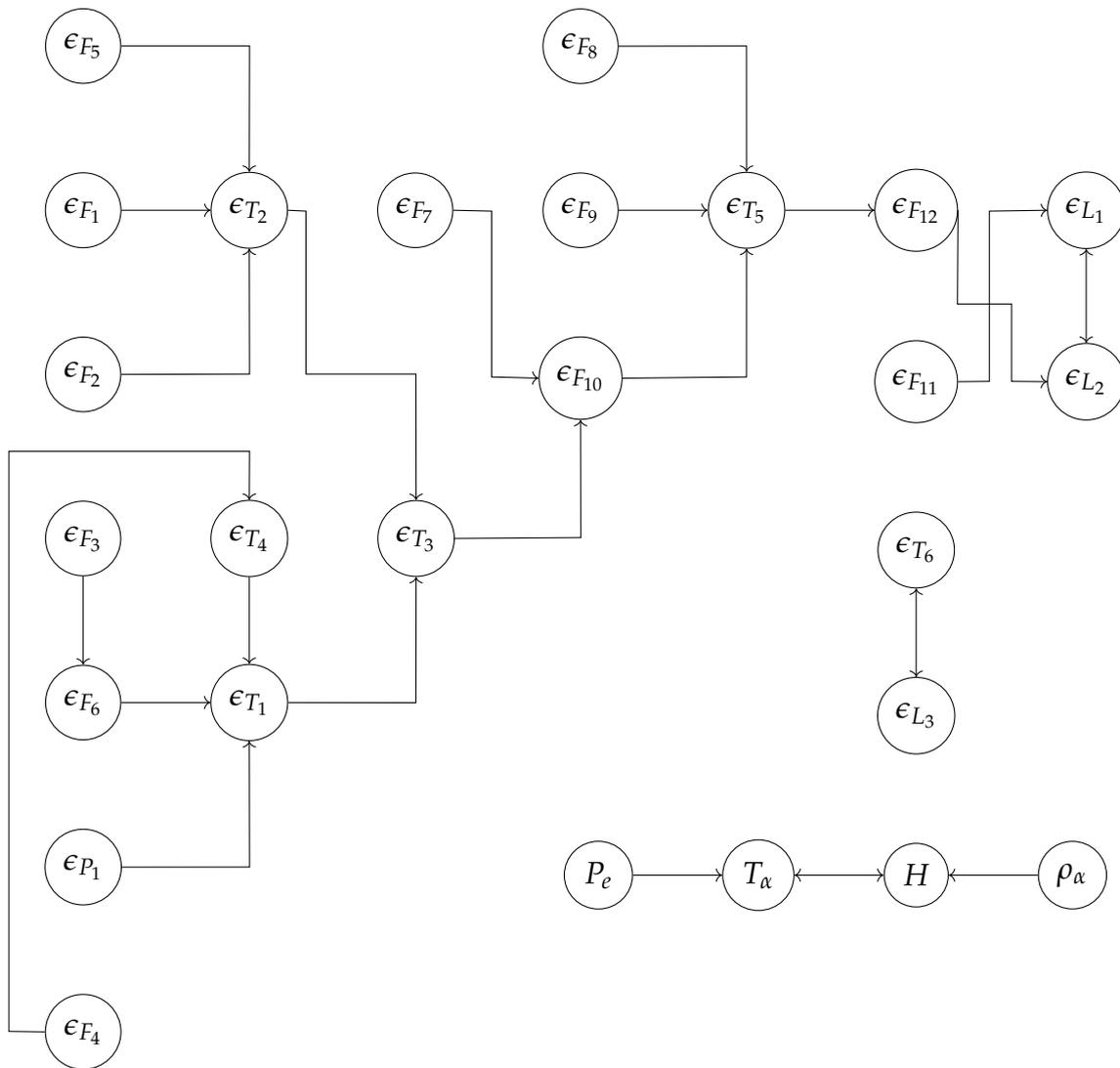


FIGURE 10.4: The causality diagram of the ammonia synthesis installation – 3rd month dataset

10.2. Validation with the ammonia synthesis installation dataset

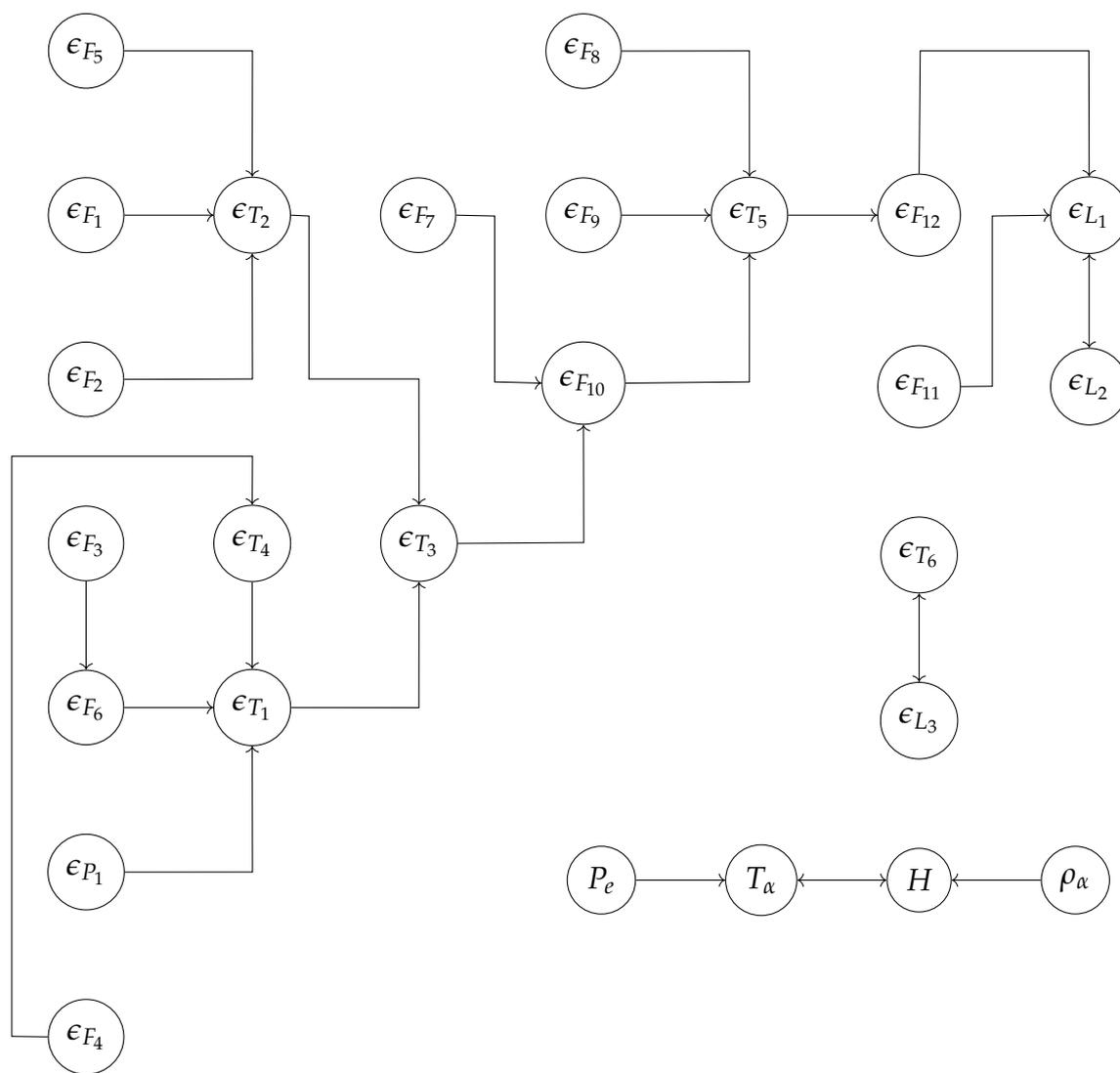


FIGURE 10.5: The causality diagram of the ammonia synthesis installation – 4th month dataset

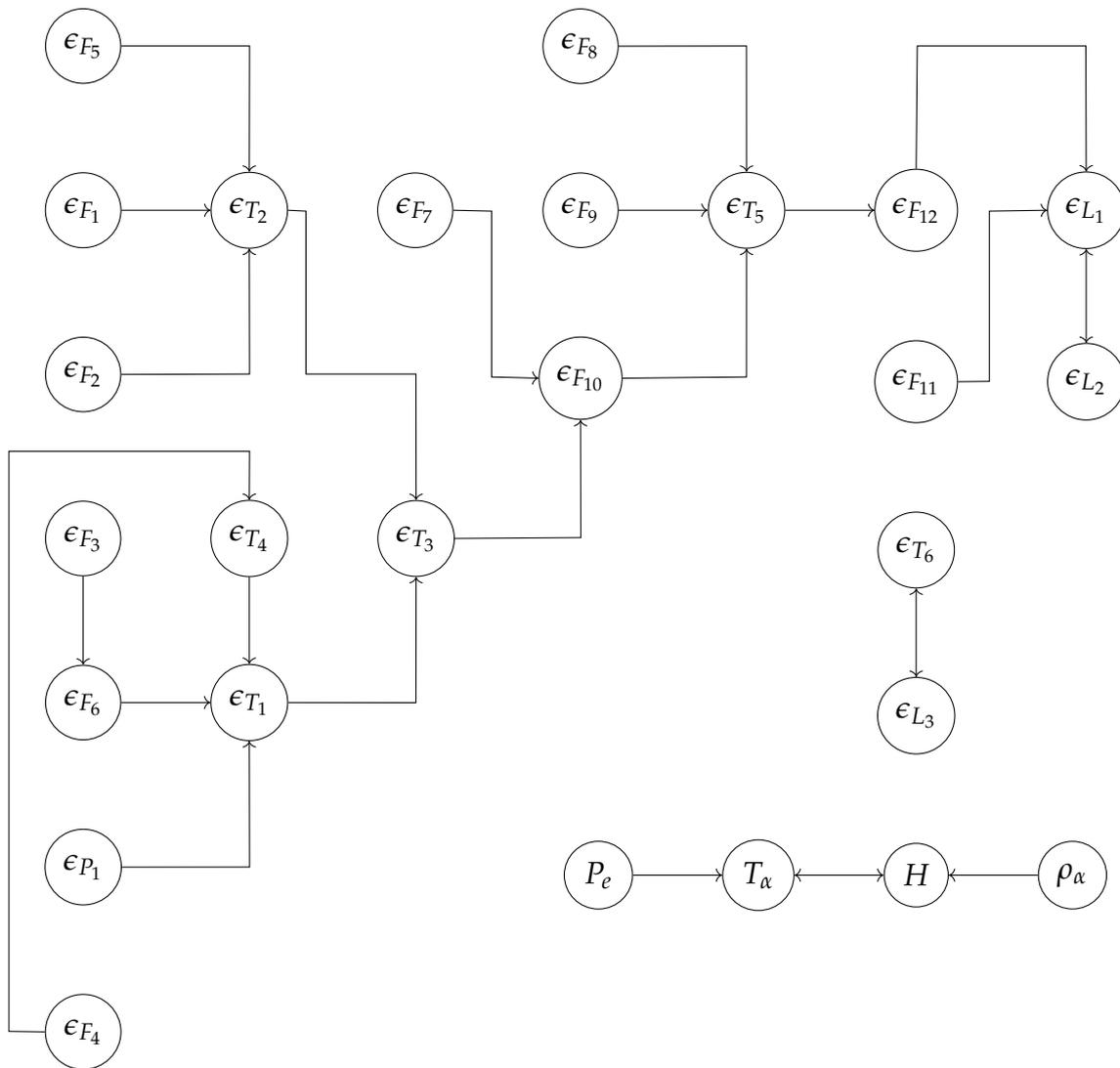


FIGURE 10.6: The causality diagram of the ammonia synthesis installation – 5th month dataset

10.2. Validation with the ammonia synthesis installation dataset

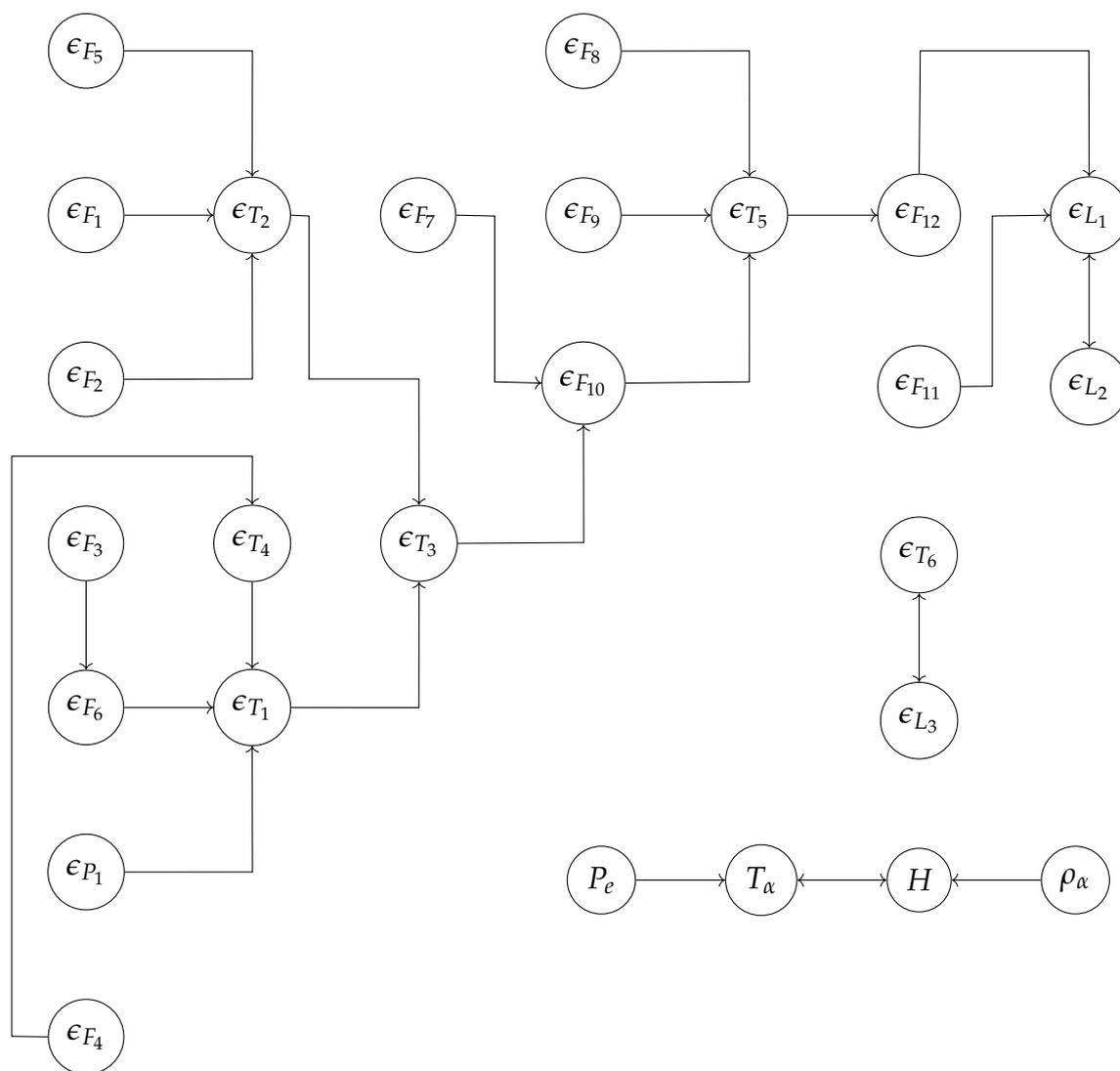


FIGURE 10.7: The causality diagram of the ammonia synthesis installation – 6th month dataset

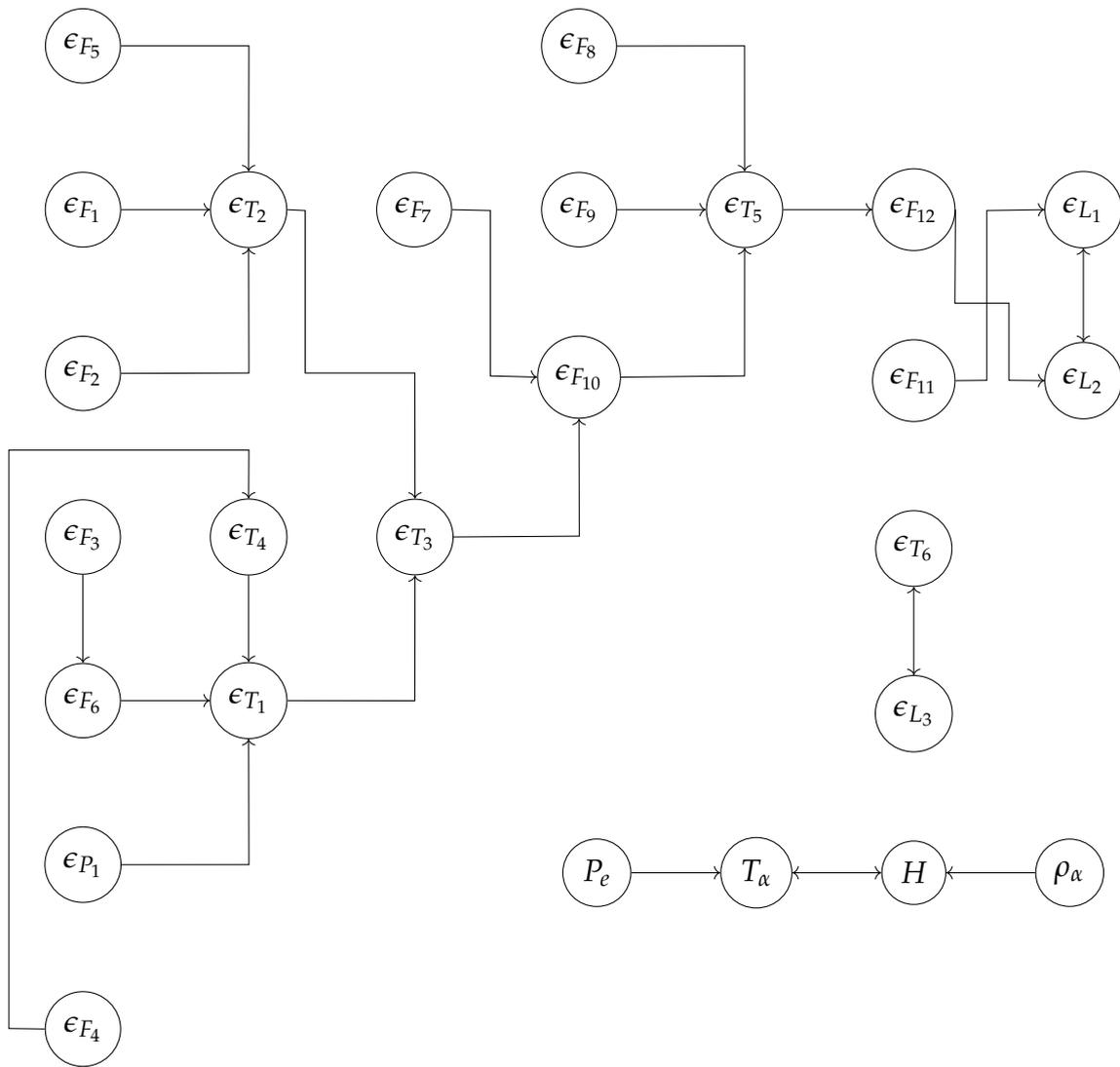


FIGURE 10.8: The causality diagram of the ammonia synthesis installation – 7th month dataset

10.2. Validation with the ammonia synthesis installation dataset

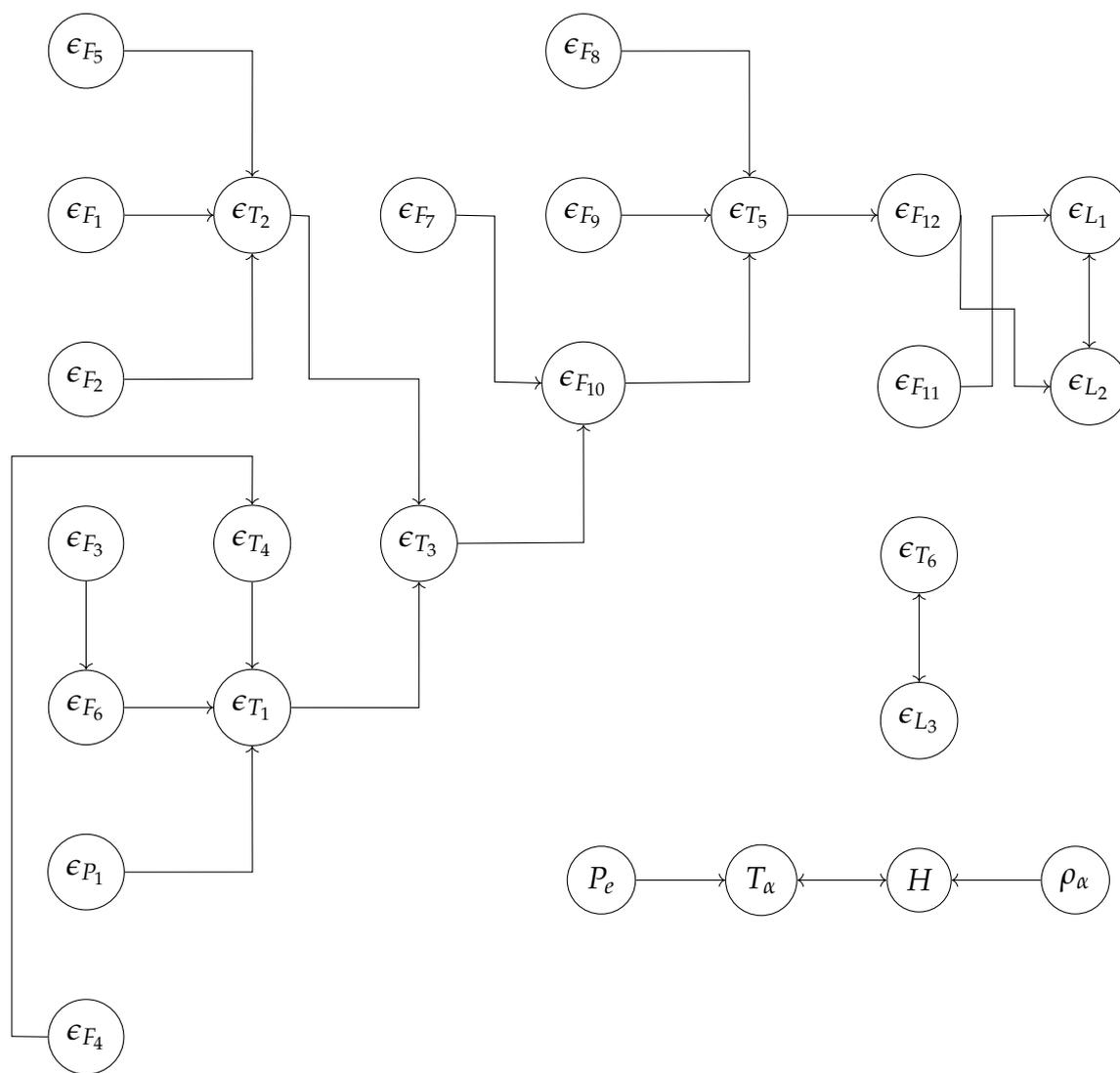


FIGURE 10.9: The causality diagram of the ammonia synthesis installation – 8th month dataset

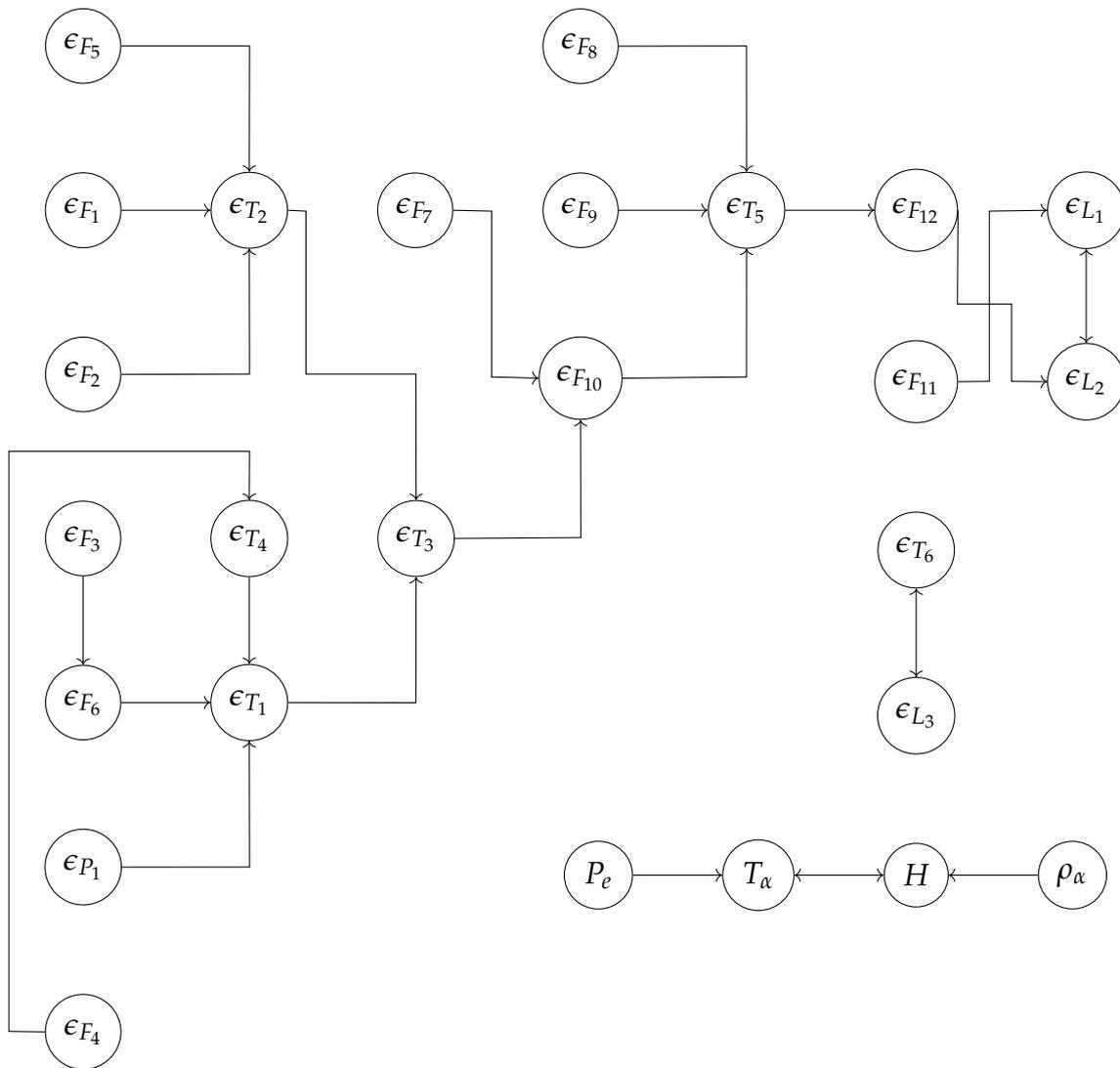


FIGURE 10.10: The causality diagram of the ammonia synthesis installation – 9th month dataset

10.2. Validation with the ammonia synthesis installation dataset

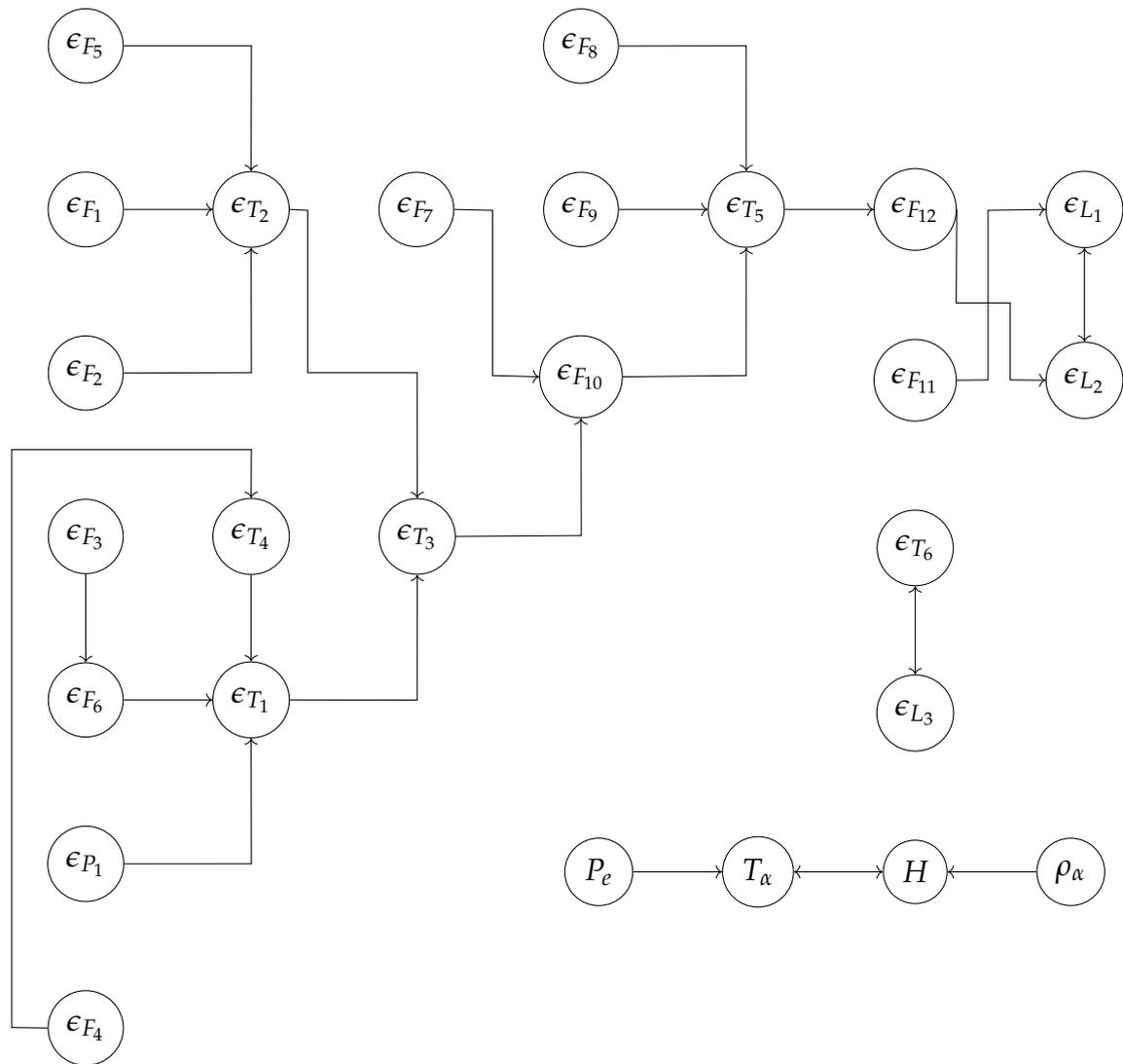


FIGURE 10.11: The causality diagram of the ammonia synthesis installation – 10th month dataset

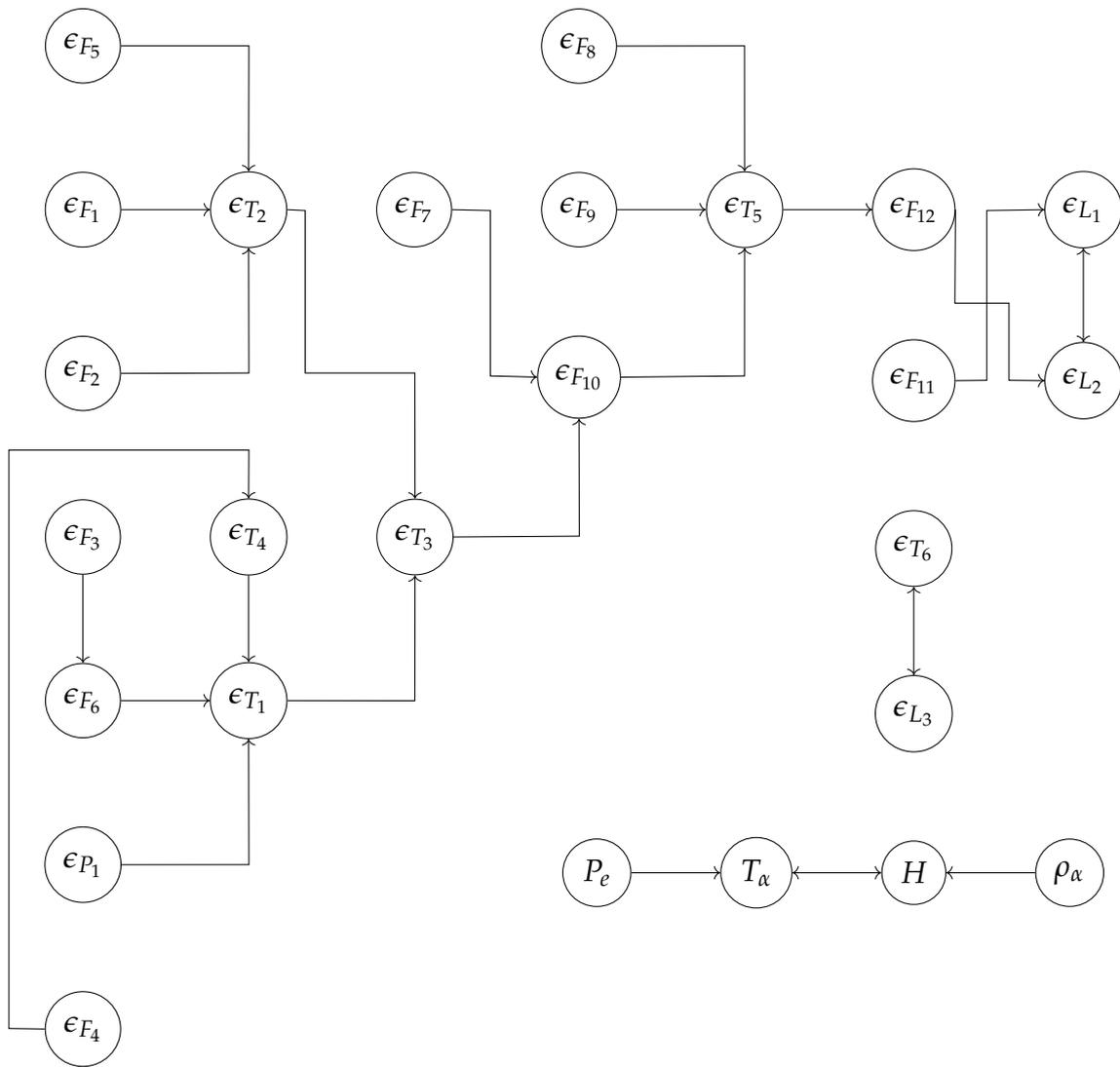


FIGURE 10.12: The causality diagram of the ammonia synthesis installation – 11th month dataset

10.2. Validation with the ammonia synthesis installation dataset

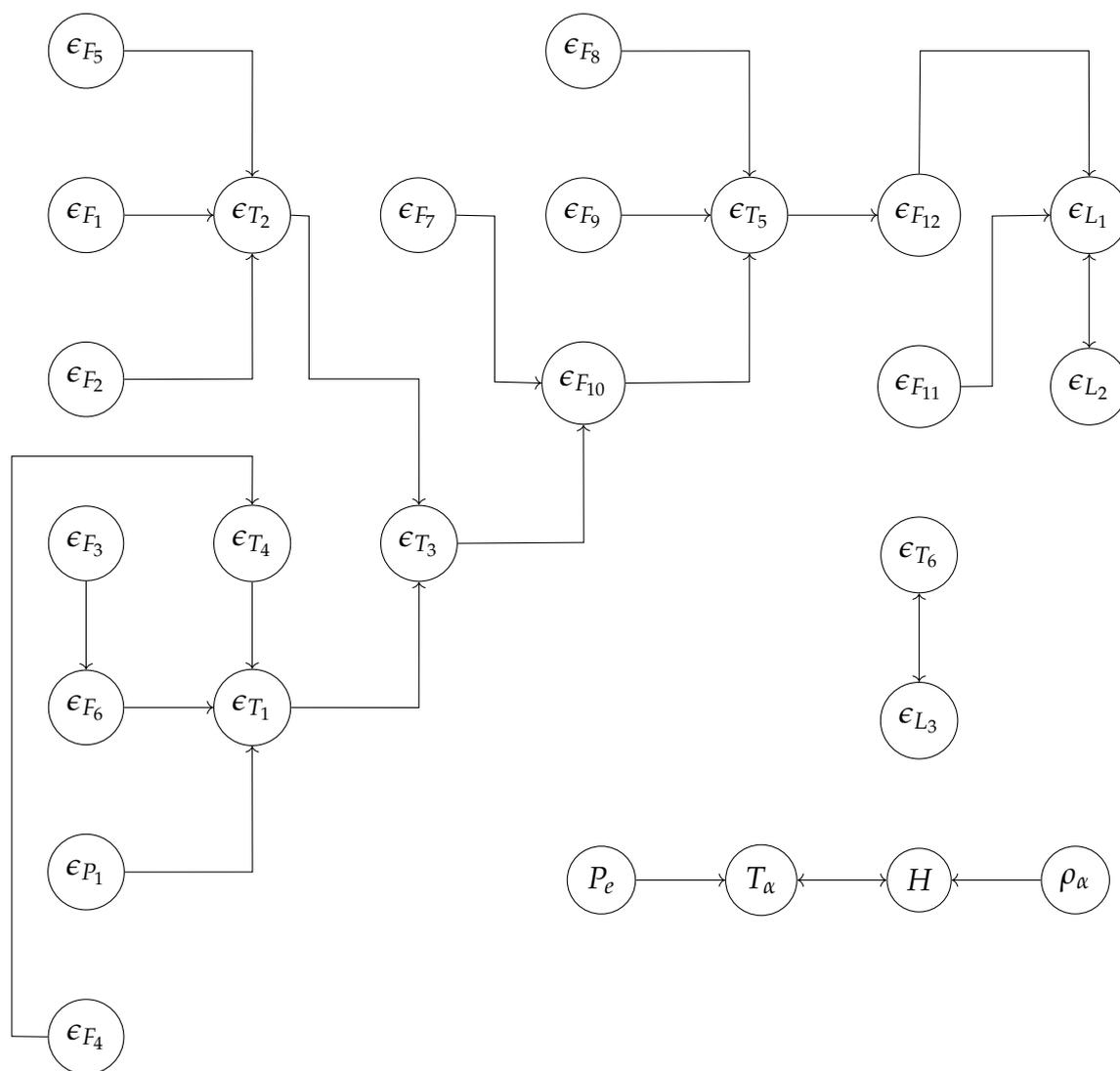


FIGURE 10.13: The causality diagram of the ammonia synthesis installation – 12th month dataset

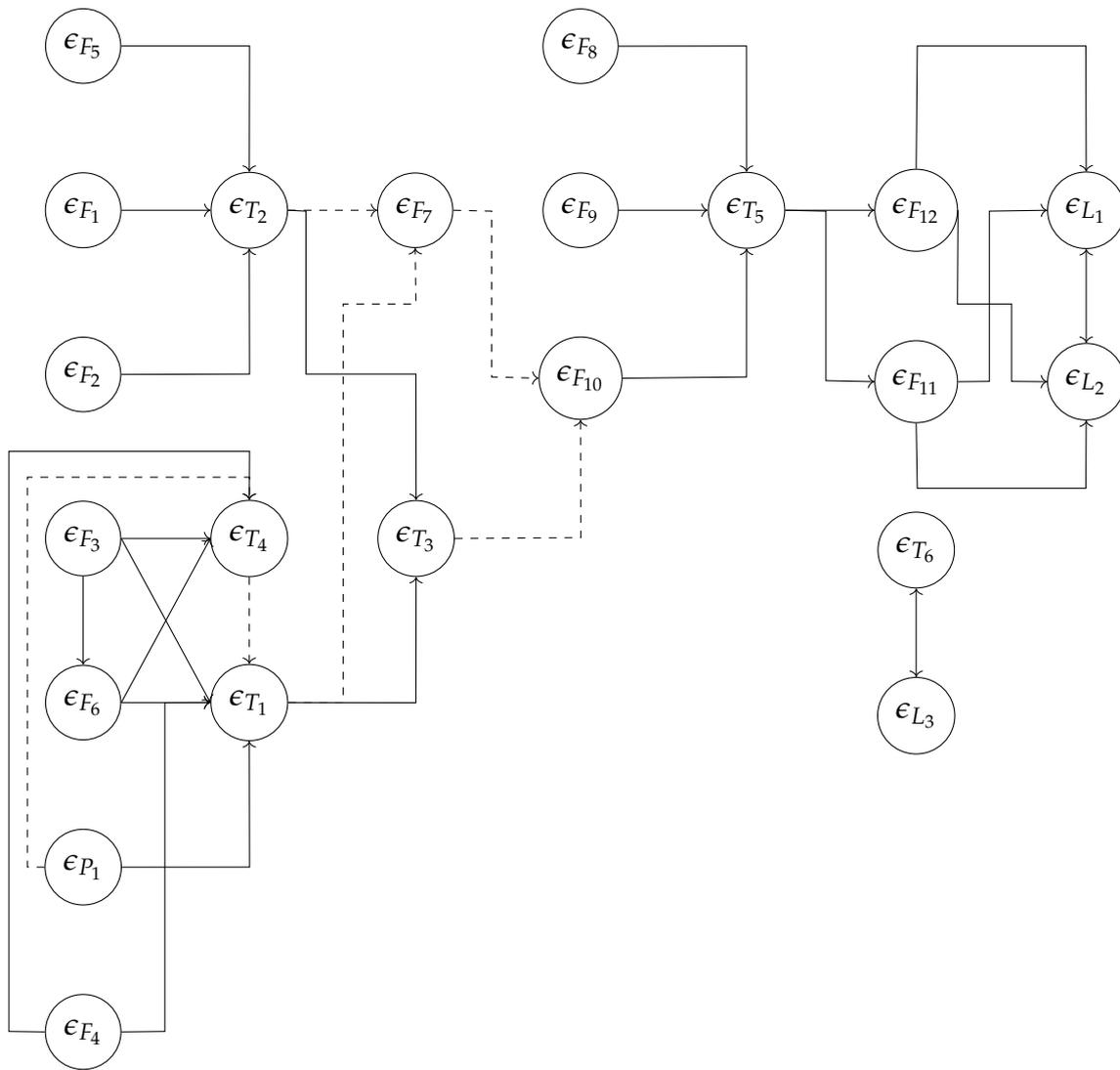


FIGURE 10.14: The causality diagram of the ammonia synthesis installation – 13th month dataset

10.2. Validation with the ammonia synthesis installation dataset

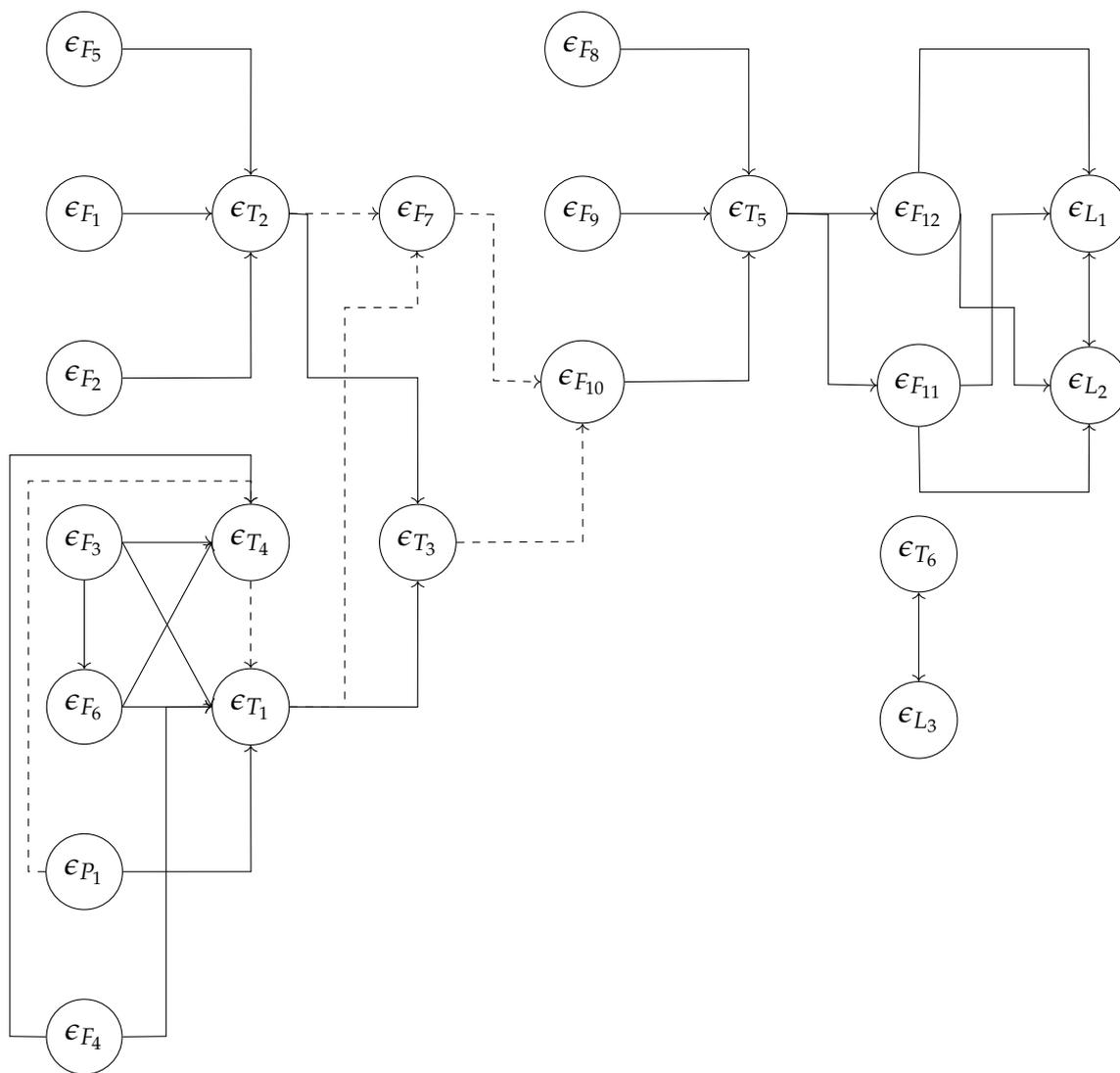


FIGURE 10.15: The causality diagram of the ammonia synthesis installation – 14th month dataset

Chapter 11

Conclusion

This thesis focuses on a comprehensive exploration of root-cause analysis in multi-loop complex control systems. It emphasizes the applicability and significance of the Transfer Entropy method as a powerful tool for causality analysis in complex systems. The performed research leads to the evaluation of the original causality assessment procedure for conducting causality analysis of the multi-loop control systems. The utilization of a simulation system allows for a rigorous examination of the method's performance and the validation of the developed analytical procedure in real-world scenarios.

The fundamental element of this research is the selection of the appropriate signal type for causality analysis. The deliberate and well-justified choice of control error as a vital information medium aligns perfectly with the principles of the Transfer Entropy approach. The consistent application of identical data processing techniques across simulation datasets, coupled with multiple iterations of the Transfer Entropy method, highlights the pivotal role played by the unique characteristics and dynamics of selected data in shaping causal outcomes. This approach not only facilitates the application of objective data analysis methodologies but also illuminates profound insights into the chosen analytical approach.

Indeed, this research underscores the remarkable versatility of the Transfer Entropy method, demonstrating its relevance across diverse domains, transcending the traditional boundaries of humanities research, and extending its utility into the domain of technical disciplines. However, it is imperative to acknowledge and address significant limitations and prerequisites vital for ensuring the method's efficacy and resilience. These constraints can be primarily categorized into two domains: those stemming from the inherent nature of the system and its dataset characteristics, and those arising from the specific Transfer Entropy properties.

One of the most notable intrinsic limitation of Transfer Entropy is the absence of an objectively determined threshold for establishing relationships among control errors. This concern is a recurrent theme in the research, as Transfer Entropy values often exhibited remarkable similarity. This observation necessitates the re-consideration of the strict criterion of selecting the highest coefficient value as the sole objective criterion. Consequently, it is assumed a singular relationship between pairs of examined control errors, thereby precluding the determination of dependencies involving one control loop and two or more others – a scenario frequently

encountered in multi-loop control systems. This simplified depiction of the system's structure may have potentially overlooked crucial relationships with significant implications for overall system behavior. An intriguing avenue for further research involves the consideration of time delay between control errors as an additional factor in the determination of the relationships between variables, potentially expanding the developed procedure.

The doubts, deliberations, and conclusions drawn from this research lead to the formulation of the original **causality assessment procedure** for conducting the root-cause analysis using the Transfer Entropy method. This procedure is successfully validated using real data from the ammonia synthesis installation, graciously provided by Grupa Azoty Zakłady Azotowe "Puławy" SA.

In conclusion, this research endeavor makes significant contributions to the understanding of the Transfer Entropy method's applicability and leads to the development of the scientifically grounded **causality assessment procedure** for the comprehensive analysis of complex multi-loop control systems. The proposed methodology, tailored to accommodate both poorly-tuned and well-tuned systems, takes into account the tuning quality, data preprocessing, distribution analysis, and advanced techniques for revealing causal relationships. The results underscore the significance of the selection of appropriate data which are used during the analysis. The observations present the potential and limitations of the Transfer Entropy method.

These findings may open new research areas, which now are rarely addressed. The potential and industrial implications of the causality analysis and the Transfer Entropy are significant. The reliable root-cause analysis introduces a new perspective to the monitoring, fault diagnosis, alarming, and maintenance of complex multi-loop control systems.

List of Figures

2.1	Implementation procedure of a conventional FDDI method	4
2.2	CPA industrial utilization process (Domański, 2020)	9
2.3	CPA techniques classification (Domański, 2020)	9
3.1	Component block diagram of a single element closed-loop system (Jelali, 2013)	15
3.2	Component block diagram of a network of single-loop systems . . .	16
3.3	The sample layout of causality diagram	25
3.4	An illustrative representation of the Gaussian normal distribution as a function of mean and standard deviation	26
3.5	An illustrative representation of the α -stable distribution as a function of stability index, skewness, location factor, and scale	29
3.6	An illustrative representation of the Cauchy distribution as a function of location parameter and scale factor	30
3.7	An illustrative representation of the Laplace double exponential distribution as a function of offset factor and scale parameter	31
3.8	An illustrative representation of the t Location-scale distribution as a function of location factor, scale parameter, and shape parameter . .	32
3.9	Parameter selection for the fixed-bins algorithm (taken from Lee et al., 2012)	33
3.10	Parameter selection for the KDE algorithm (taken from Lee et al., 2012)	34
4.1	Simulation environment presenting multi-loop PID-based control layout (Falkowski and Domański, 2023)	40
4.2	Layout of the industrial realization of the feedforward disturbance decoupling	41
4.3	Control errors time series for the simulation data including Gaussian noise	42
4.4	Control errors time series for the simulation data including Gaussian noise and Cauchy disturbance	43
4.5	The actual causality diagram of the simulated benchmark	44
4.6	Control error ϵ_1 histogram for the simulation data including Gaussian noise with PDF fitting	44
4.7	Control error ϵ_2 histogram for the simulation data including Gaussian noise with PDF fitting	45

4.8	Control error ϵ_3 histogram for the simulation data including Gaussian noise with PDF fitting	45
4.9	Control error ϵ_4 histogram for the simulation data including Gaussian noise with PDF fitting	45
4.10	Control error ϵ_5 histogram for the simulation data including Gaussian noise with PDF fitting	46
4.11	Control error ϵ_1 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting	46
4.12	Control error ϵ_2 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting	47
4.13	Control error ϵ_3 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting	47
4.14	Control error ϵ_4 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting	47
4.15	Control error ϵ_5 histogram for the simulation data including Gaussian noise and Cauchy disturbance with PDF fitting	48
4.16	Ammonia production plant layout (Domański et al., 2023)	49
4.17	Selected control errors of flow, level, pressure, and temperature of the ammonia synthesis installation - dataset for the 8 th month of operation	51
4.18	Selected weather parameters of ambient air temperature, air density, humidity level, and atmospheric pressure of the ammonia synthesis installation - dataset for the 8 th month of operation	52
4.19	Probability distribution of the flow control error (ϵ_{F_7}) over a 14-month operational period of the ammonia synthesis installation	53
4.20	Probability distribution of the level control error (ϵ_{L_1}) over a 14-month operational period of the ammonia synthesis installation	53
4.21	Probability distribution of the pressure control error (ϵ_{P_1}) over a 14-month operational period of the ammonia synthesis installation	54
4.22	Probability distribution of the temperature control error (ϵ_{T_2}) over a 14-month operational period of the ammonia synthesis installation	54
4.23	The actual causality diagram of the ammonia synthesis installation	64
5.1	Control error ϵ_1 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting	66
5.2	Control error ϵ_2 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting	66
5.3	Control error ϵ_3 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting	67
5.4	Control error ϵ_4 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting	67
5.5	Control error ϵ_5 histogram for the simulation data including Gaussian noise with Gaussian distribution fitting	67
5.6	Control error ϵ_1 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting	68

List of Figures

5.7	Control error ϵ_2 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting . . .	68
5.8	Control error ϵ_3 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting . . .	69
5.9	Control error ϵ_4 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting . . .	69
5.10	Control error ϵ_5 histogram for the simulation data including Gaussian noise and Cauchy disturbance with Gaussian distribution fitting . . .	70
5.11	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance	73
	a The actual causality diagram of the simulated benchmark . . .	73
	b The obtained causality diagram of the simulated benchmark . .	73
6.1	Control error ϵ_1 histogram for the simulation data including Gaussian noise with selected PDF fitting	75
6.2	Control error ϵ_2 histogram for the simulation data including Gaussian noise with selected PDF fitting	76
6.3	Control error ϵ_3 histogram for the simulation data including Gaussian noise with selected PDF fitting	76
6.4	Control error ϵ_4 histogram for the simulation data including Gaussian noise with selected PDF fitting	76
6.5	Control error ϵ_5 histogram for the simulation data including Gaussian noise with selected PDF fitting	77
6.6	Control error ϵ_1 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting	77
6.7	Control error ϵ_2 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting	77
6.8	Control error ϵ_3 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting	78
6.9	Control error ϵ_4 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting	78
6.10	Control error ϵ_5 histogram for the simulation data including Gaussian noise and Cauchy disturbance with selected PDF fitting	78
6.11	Comparison of the actual and obtained causality graphs - data impacted by Gaussian noise	81
	a The actual causality diagram of the simulated benchmark . . .	81
	b The causality diagram of the simulated benchmark – dataset with Gaussian noise	81
6.12	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance	82
	a The actual causality diagram of the simulated benchmark . . .	82
	b The causality diagram of the simulated benchmark – dataset with Gaussian noise and Cauchy disturbance	82

6.13	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise	84
a	The actual causality diagram of the simulated benchmark	84
b	The causality diagram of the simulated benchmark - dataset with Gaussian noise	84
6.14	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise	85
a	The actual causality diagram of the simulated benchmark	85
b	The causality diagram of the simulated benchmark - dataset with Gaussian noise	85
6.15	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise	86
a	The actual causality diagram of the simulated benchmark	86
b	The causality diagram of the simulated benchmark – dataset with Gaussian noise	86
6.16	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance	87
a	The actual causality diagram of the simulated benchmark	87
b	The causality diagram of the simulated benchmark – dataset with Gaussian noise and Cauchy disturbance	87
6.17	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance	88
a	The actual causality diagram of the simulated benchmark	88
b	The causality diagram of the simulated benchmark – dataset with Gaussian noise and Cauchy disturbance	88
6.18	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance	90
a	The actual causality diagram of the simulated benchmark	90
b	The causality diagram of the simulated benchmark – dataset with Gaussian noise and Cauchy disturbance	90
7.1	Results of trend identification and its removal from ϵ_1 using spline interpolation – dataset with Gaussian noise and Cauchy disturbance	93
7.2	Results of trend identification and its removal from ϵ_2 using spline interpolation – dataset with Gaussian noise and Cauchy disturbance	93
7.3	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance	96
a	The actual causality diagram of the simulated benchmark	96
b	The causality diagram of the simulated benchmark – dataset with Gaussian noise and Cauchy disturbance	96
7.4	Comparison of the actual and obtained causality graphs - data impacted by Gaussian noise and Cauchy disturbance	98
a	The actual causality diagram of the simulated benchmark	98

List of Figures

b	The causality diagram of the simulated benchmark – dataset with Gaussian noise and Cauchy disturbance	98
8.1	Results of decomposition process using MEEMD algorithm for control error ϵ_1 – dataset with Gaussian noise	100
8.2	Results of decomposition process using MEEMD algorithm for ϵ_1 – dataset with Gaussian noise and Cauchy disturbance	101
8.3	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise	103
a	The actual causality diagram of the simulated benchmark	103
b	The causality diagram of the simulated benchmark - dataset with Gaussian noise	103
8.4	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance	105
a	The actual causality diagram of the simulated benchmark	105
b	The causality diagram of the simulated benchmark - dataset with Gaussian noise and Cauchy disturbance	105
9.1	Oscillatory signals obtained in the process of MEEMD decomposition for control error ϵ_1 – dataset with Gaussian noise	108
9.2	Oscillatory signals obtained in the process of MEEMD decomposition for ϵ_1 – dataset with Gaussian noise and Cauchy disturbance	109
9.3	Comparison of the actual and obtained causality graphs – data impacted by Gaussian noise and Cauchy disturbance	113
a	The actual causality diagram of the simulated benchmark	113
b	The causality diagram of the simulated benchmark - dataset with Gaussian noise and Cauchy disturbance	113
10.1	Diagram of the proposed analytical procedure for causality analysis using the Transfer Entropy approach	119
10.2	The causality diagram of the ammonia synthesis installation – 1 st month dataset	122
10.3	The causality diagram of the ammonia synthesis installation – 2 nd month dataset	123
10.4	The causality diagram of the ammonia synthesis installation – 3 rd month dataset	124
10.5	The causality diagram of the ammonia synthesis installation – 4 th month dataset	125
10.6	The causality diagram of the ammonia synthesis installation – 5 th month dataset	126
10.7	The causality diagram of the ammonia synthesis installation – 6 th month dataset	127
10.8	The causality diagram of the ammonia synthesis installation – 7 th month dataset	128

List of Figures

10.9	The causality diagram of the ammonia synthesis installation – 8 th month dataset	129
10.10	The causality diagram of the ammonia synthesis installation – 9 th month dataset	130
10.11	The causality diagram of the ammonia synthesis installation – 10 th month dataset	131
10.12	The causality diagram of the ammonia synthesis installation – 11 th month dataset	132
10.13	The causality diagram of the ammonia synthesis installation – 12 th month dataset	133
10.14	The causality diagram of the ammonia synthesis installation – 13 th month dataset	134
10.15	The causality diagram of the ammonia synthesis installation – 14 th month dataset	135

List of Tables

2.1	Systematic categorization of data-driven FDDI methods	5
2.2	Methods classification based on different applications	11
4.1	Tuning parameters of simulated controllers R_i – Matlab PID implementation in a pararel formulation	40
4.2	Gaussian noise and Cauchy disturbance signals parameters	44
4.3	Gaussian statistical metric (mean) of the ammonia synthesis installation control errors	57
4.4	Gaussian statistical metric (median) of the ammonia synthesis installation control errors	58
4.5	Gaussian statistical metric (standard deviation) of the ammonia synthesis installation control errors	59
4.6	Gaussian statistical metric (kurtosis) of the ammonia synthesis installation control errors	60
4.7	Gaussian statistical metric (skewness) of the ammonia synthesis installations control errors	61
4.8	Gaussian statistical metric (mean absolute deviation) of the ammonia synthesis installation control errors	62
5.1	Parameters of Transfer Entropy function for the simulation data . . .	70
5.2	Calculated Transfer Entropy coefficients for the simulation data – Gaussian noise	71
5.3	Calculated Transfer Entropy coefficients for the simulation data – Gaussian noise and Cauchy disturbance	72
6.1	Fit indicators for the dataset with Gaussian noise	79
6.2	Fit indicators for the dataset with Gaussian noise and Cauchy disturbance	79
6.3	Calculated Transfer Entropy coefficients based on α -stable distribution for the dataset with Gaussian noise	80
6.4	Calculated Transfer Entropy coefficients based on α -stable distribution for the dataset with Gaussian noise and Cauchy disturbance . .	82
6.5	Calculated Transfer Entropy coefficients based on Darbellay-Vajda algorithm for the dataset with Gaussian noise	83
6.6	Calculated Transfer Entropy coefficients based on Fixed Bins algorithm for the dataset with Gaussian noise	84

6.7	Calculated Transfer Entropy coefficients based on Kernel Density Estimation method for the dataset with Gaussian noise	85
6.8	Calculated Transfer Entropy coefficients based on Darbellay-Vajda algorithm for the dataset with Gaussian noise and Cauchy disturbance	87
6.9	Calculated Transfer Entropy coefficients based on Fixed Bins algorithm for the dataset with Gaussian noise and Cauchy disturbance	88
6.10	Calculated Transfer Entropy coefficients based on Kernel Density Estimation method for the dataset with Gaussian noise and Cauchy disturbance	89
7.1	Mean absolute deviation (MAD) values for i^{th} polynomial order - dataset with Gaussian noise	91
7.2	Mean absolute deviation (MAD) values for i^{th} polynomial order - dataset with Gaussian noise and Cauchy disturbance	92
7.3	Calculated thresholds values using ESD, IQR and Hampel filter - dataset with Gaussian noise	94
7.4	Calculated thresholds values using ESD, IQR and Hampel filter - dataset with Gaussian noise and Cauchy disturbance	94
7.5	Calculated Transfer Entropy coefficients using IQR outliers detection method for the dataset with Gaussian noise	95
7.6	Calculated Transfer Entropy coefficients using IQR outliers detection method for the dataset with Gaussian noise and Cauchy disturbance	97
8.1	Calculated Transfer Entropy coefficients for noise signals - dataset with Gaussian noise	102
8.2	Calculated Transfer Entropy coefficients for noise signals - dataset with Gaussian noise and Cauchy disturbance	104
9.1	Calculated Transfer Entropy coefficients for oscillatory signals - dataset with Gaussian noise	111
9.2	Calculated Transfer Entropy coefficients for oscillatory signals - dataset with Gaussian noise and Cauchy disturbance	112

List of Abbreviations

AMP	A mplitude Index
ANNs	A rtificial n eural n etworks
AR	A utoregressive
Cc	C ross-correlation
CoE	C auses of E ffects
CPA	C ontrol P erformance A ssessment
CV	P rocess variable
DNN	D ynamic n eural n etwork
DPCA	D ynamic P rincipal C omponent analysis
DPLS	D iscriminant p artial l east s quare
DV	D arbellay- V ajda
EEMD	E nsemble E mpirical M ode D ecomposition
EMD	E mpirical M ode D ecomposition
EoC	E ffects of C auses
ESD	E xtrême S tudentized D eviate
EWMA	E xponentially W eighted M oving A verages
FB	F ixed- B ins
FDA	F isher's d iscriminant analysis
FDDI	F ault D etection D iagnosis and I solation
GC	G ranger C ausality
GMM	G aussian m ixture m odel
HMM	H idden M arkov m odel
HSMM	H idden s emi- M arkov m odel
IAE	I ntegral A bsolute E rror
ICA	I ndependent C omponent analysis
IMF	I ntrinsic M ode F unction
Inf	I nfinity
IQR	I nter Q uartile R ange
ISTC	I ntegral of S quare T ime derivative of the C ontrol input
ITAE	I ntegral T ime A bsolute V alue
KDE	K ernel D ensity E stimation
KDE	K ernel D ensity E stimation
KICA	K ernel I ndependent C omponent analysis
KPCA	K ernel P rincipal C omponent analysis
KPIs	K ey P erformance I ndicators
KPLS	K ernel p artial l east s quares

List of Abbreviations

LKPCA	Local Kernel Principal Component analysis
M-estimator	Maximum Likelihood Estimator
MAD	Mean Absolute Deviation
MCS	Monte Carlo simulation
MDist	Modified Distance
MEMD	Median Ensemble Empirical Mode Decomposition
MM	Markov model
MS	Mode Splitting
MSE	Mean Square Error
MV	Manipulated variable
NA	Not Applicable
NaN	Not a Number
OP	Output signal
PCA	Principal Component analysis
PDC	Partial Directed Coherence
PDF	Probabilistic Density Function
PF	Particle-filtering
PID	Proportional-Integral-Derivative
PLS	Partial least squares
PM-FD	Process monitoring and fault diagnosis
RPCA	Recursive Principal Component analysis
RPLS	Recursive partial least squares
SP	Set point
SVM	Support vector machine
TE	Transfer Entropy
TSV	Total Squared Variation
WSS	Wide-Sense-Stationary

Physical Constants

Ammonia	NH_3
Carbon Dioxide	CO_2
Carbon Monoxide	CO
Hydrogen	H_2
Methane	CH_4
Nitrogen	N_2
Oxygen	O_2
Water	H_2O
Hertz	Hz
Pressure	MPa

List of Symbols

C_{thumb}	Rule of thumb coefficient
$d(t)$	Intrinsic Mode Function
$F(s)$	Feedforward filter transfer function
$F_d(s)$	Filter to shape the fat-tailed disturbance transfer function
$F_{x \rightarrow y}^{CC}$	Cross-correlation coefficient
$F_{\delta, \gamma}^{Cauchy}(x)$	Cauchy probabilistic density function
$F_{\mu, \sigma}^{Gauss}(x)$	Normal probability distribution function
$F_{x \rightarrow y}^{GC}$	Granger Causality coefficient
$F_{\delta, \gamma}^{Lap}(x)$	Laplace double exponential distribution function
$F_{\alpha, \beta, \delta, \gamma}^{stab}(x)$	α -stable distribution function
$F_{\delta, \gamma, \nu}^{t-LS}(x)$	t Location-scale distribution function
$G(s)$	Simulated process transfer function
G_d	Simulated process transfer function
h	Prediction horizon
H	Air humidity level
I	Multi-component signal
k	Number of process variables
K	Gain
l	Sampling interval
N	Number of samples
N_{IMF}	Number of Intrinsic Mode Functions
N_o	Outliers number
N_p	Number of data points
p	Complete / conditional Probability Density Function
P_e	External atmospheric pressure
q	Model order
Q	Quantization level
Q_1	First quartile (lower)
Q_3	Third quartile (upper)
R	PID / PI controller
$r(t)$	Setpoint
$Res_{N_{IMF}}$	Residue of N_{IMF} intrinsic modes
T	Time constant
T_α	Ambient air temperature

$T_{x \rightarrow y}$	Transfer Entropy coefficient
$u(t)$	Controller output
$u_v(t)$	Manipulated variable
v	Noise amplitude
$x(t)$	Process variable x
$y(t)$	Process variable y
$z(t)$	Output
\mathbf{B}	AR polynomial coefficients
\mathbf{X}	Vector of x process variables
\mathbf{Y}	Vector of y process variables
α	Stability index
β	Skewness parameter
κ	Scale M-estimator
Σ	Noise covariance matrix
δ	Distribution location
$\epsilon(t)$	Control error
ϵ_F	Control error of flow
ϵ_L	Control error of level
ϵ_P	Control error of pressure
ϵ_T	Control error of temperature
η	Kurtosis parameter
γ	Distribution scale factor
γ_C	Cauchy disturbance
$\hat{\epsilon}(t)$	Prediction error
λ	Multiplier for scaling in Probability Density Function
μ	Mean
ν	Shape function
ρ_α	Air density
σ	Standard deviation
σ^2	Variance
σ_{cov}	Covariance
σ_G	Gaussian noise
τ	Time delay (lag)
$\tilde{\mu}$	Median
$ \hat{\pi}(\omega) $	Partial Directed Coherence coefficient
$\mathbf{Y}(t)$	Vector of multivariate manipulated variable noise terms
$v(t)$	Manipulated variable noise terms
$\Gamma(\cdot)$	Gamma function
ψ_{log}	Logistic scale estimator

Bibliography

- al., Cross-Disorder Group of the Psychiatric Genomics Consortium et (2013). "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis". In: *The Lancet* 381.9875, pp. 1371–1379.
- Alagoz, B. B. et al. (2015). "Implicit disturbance rejection performance analysis of closed loop control systems according to communication channel limitations". In: *IET Control Theory Applications* 9.17, pp. 2522–2531.
- Ancona, N., D. Marinazzo, and S. Stramaglia (2004). "Radial basis function approach to nonlinear Granger causality of time series". In: *Phys. Rev. E* 70.
- Åström, K. J. (1967). "Computer control of a paper machine - an application of linear stochastic control theory". In: *IBM Journal* 11, pp. 389–405.
- Åström, K. J. and T. Hägglund (2000). "Benchmark Systems for PID Control". In: *IFAC Digital Control: Past, Present and Future of PID Control*, pp. 165–166.
- Astrom, K. J. et al. (2001). "Control of Complex Systems". In: *Springer-Verlag, London*.
- Baccala, L.A. and K. Sameshima (2001). "Partial directed coherence: a new concept in neural structure determination". In: *Biological cybernetics* 84.6, pp. 463–474.
- Bauer, M. and N.F. Thornhill (2008). "A practical method for identifying the propagation path of plant-wide disturbances". In: *Journal of Process Control* 18.78, pp. 707–719.
- Bauer, M. et al. (2007). "Finding the direction of disturbance propagation in a chemical process using transfer entropy". In: *IEEE Transactions on Control Systems Technology* 15.1, pp. 12–21.
- Bauer, M. et al. (2016). "The current state of control loop performance monitoring – A survey of application in industry". In: *Journal of Process Control* 38, pp. 1–10.
- Bequette, B. (2002). *Process Control: Modeling, Design, and Simulation*. First. Upper Saddle River, NJ, USA: Prentice Hall Press.
- Betlem, B. and B. Roffel (2003). *Advanced Practical Process Control (Advances In Soft Computing)*. Springer.
- Boccaletti, S. et al. (2002). "The synchronization of chaotic systems". In: *Phys. Rep.* 366, pp. 1–101.
- Braun, S.G. et al. (June 2002). "Encyclopedia of Vibration: Volumes 1, 2, and 3". In: *Applied Mechanics Reviews* 55.3, B45–B45. ISSN: 0003-6900. DOI: 10.1115/1.1470670. URL: <https://doi.org/10.1115/1.1470670>.
- Bressler, S.L. and A.K. Seth (2010). "Wiener- causality: a well established methodology". In: *NeuroImage* 58, pp. 323–9.

- Butte, A. J. and I. S. Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements". In: *Pac. Symp. Biocomput.*, pp. 418–429.
- Chavez, M., J. Martinerie, and M. Le Van Quyen (2003). "Statistical assessment of nonlinear causality: application to epileptic EEG signals". In: *J. Neurosci. Methods* 124, pp. 113–128.
- Chen, Y., S. L. Bressler, and M. Ding (2009). "Dynamics on networks: assessing functional connectivity with Granger causality". In: *Comput Math Organ Theory* 15, pp. 329–350.
- Chen, Y. et al. (2004). "Analyzing multiple nonlinear time series with extended Granger causality". In: *Phys. Lett. A* 324, pp. 26–35.
- Choudhury, M. A. A. S., S. L. Shah, and N. F. Thornhill (2004). "Diagnosis of poor control-loop performance using higher-order statistics". In: *Automatica* 40.10, pp. 1719–1728.
- Clements, Michael P. and David. F. Hendry (2001). "Forecasting with difference-stationary and trend-stationary models". In: *The Econometrics Journal* 4.1, S1–S19. ISSN: 13684221, 1368423X. (Visited on 05/08/2023).
- Cochran, J. M. (2017). *Introduction to Mathematical Modeling*. CreateSpace Independent Publishing Platform.
- Coles, S.G. and M.J. Dixon (1999). "Likelihood-Based Inference for Extreme Value Models". In: *Extremes* 2, pp. 5–23. DOI: 10.1023/A:1009905222644.
- Croux, C. and C. Dehon (2014). "Robust Estimation of Location and Scale". Wiley StatsRef: Statistics Reference Online.
- D'Agostino, Ralph B and Michael A Stephens (1986). *Goodness-of-Fit Techniques*. USA: Marcel Dekker, Inc. ISBN: 0824774876.
- Darbellay, G. A. and I. Vajda (1999). "Estimation of the information by an adaptive partitioning of the observation space". In: *IEEE Transactions on Information Theory* 45.4, pp. 1315–1321.
- Dehejia, R. H. and S. Wahba (1999). "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs". In: *J. Amer. Statist. Assoc.* 94.448, pp. 1053–1062.
- Ding, S. et al. (2012). "A novel scheme for key performance indicator prediction and diagnosis with application to an industrial hot strip mill". In: *IEEE Transactions on Industrial Informatics* 9.4, pp. 2239–2247.
- Ding, S. et al. (2013). "An integrated design framework of fault tolerant wireless networked control systems for industrial automatic control applications". In: *IEEE Transactions on Industrial Informatics* 9.1, pp. 462–471.
- Ding, S. X. (2014). "Data-Driven Design of Fault Diagnosis and Fault-Tolerant Control Systems". In: *Springer: London, UK*.
- Ding, S. X. et al. (2000). "A unified approach to the optimization of fault detection systems". In: *International Journal of Adaptive Control and Signal Processing* 14.7, pp. 725–745.

Bibliography

- Domański, P. D. (2015). “Non-Gaussian properties of the real industrial control error in SISO loops”. In: *Proceedings of the 19th International Conference on System Theory, Control and Computing*, pp. 877–882.
- Domański, P. D. (2016). “Non-Gaussian and persistence measures for control loop quality assessment”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26.4, p. 043105.
- (2017). “Non-Gaussian Statistical Measures of Control Performance”. In: *Control & Cybernetics* 46.3, pp. 259–290.
- (2019). “Control quality assessment using fractal persistence measures”. In: *ISA Transactions*. in press. DOI: <https://doi.org/10.1016/j.isatra.2019.01.008>.
- (2020). *Control Performance Assessment: Theoretical Analyses and Industrial Practice*. Cham: Springer International Publishing.
- Domański, P. D. (2020). “Study on Statistical Outlier Detection and Labelling”. In: *International Journal of Automation and Computing* 17.6, pp. 788–811.
- Domański, P. D. and M. Gintrowski (2017). “Alternative approaches to the prediction of electricity prices”. In: *International Journal of Energy Sector Management* 11.1, pp. 3–27.
- Domański, P. D. et al. (2018). “Robust and Asymmetric Assessment of the Benefits from Improved Control – Industrial Validation”. In: *IFAC-PapersOnLine* 51.18. 10th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2018, pp. 815–820.
- Domański, P. D. et al. (2019). “Multi-criteria Loop Quality Assessment: A Large-Scale Industrial Case Study”. In: *Proceedings of IEEE International Conference on Methods and Models in Automation and Robotics MMAR*. Miedzyzdroje, Poland, pp. 99–104.
- Domański, Paweł D. (2020). “Performance Assessment of Predictive Control—A Survey”. In: *Algorithms* 13.4. ISSN: 1999-4893. DOI: 10.3390/a13040097. URL: <https://www.mdpi.com/1999-4893/13/4/97>.
- Domański, Paweł D. et al. (2023). “Assessing Control Sustainability Using L-Moment Ratio Diagrams”. In: *Electronics* 12.11. ISSN: 2079-9292. DOI: 10.3390/electronics12112377. URL: <https://www.mdpi.com/2079-9292/12/11/2377>.
- Domanski, P.D. (2020). “Control Performance Assessment: Theoretical Analyses and Industrial Practice”. In: *Springer International Publishing, Cham*.
- Dong, H., Z. Wang, and H. Gao (2012a). “Fault detection for Markovian jump systems with sensor saturations and randomly varying nonlinearities”. In: *IEEE Transactions on Circuits and Systems I* 59.10, pp. 2354–2362.
- (2012b). “On design of quantized fault detection filters with randomly occurring nonlinearities and mixed time-delays”. In: *Signal Processing* 92.4, pp. 1117–1125.
- Dong, H. et al. (2011). “Robust H_∞ filtering for Markovian jump systems with randomly occurring nonlinearities and sensor saturation: the finite-horizon case”. In: *IEEE Transactions on Signal Processing* 59.7, pp. 3048–3057.
- Duan, P. (2014). “Information Theory-based Approaches for Causality Analysis with Industrial Applications”. In: *Thesis, University of Alberta*.

- Dziuba, K. et al. (2018). "Multicriteria control quality assessment for ammonia production process (in Polish)". In: *1st Scientific and Technical Conference Innovations in the Chemical Industry*. Ed. by A. Zalewska. Warszawa, Poland, pp. 80–90.
- Dziuba, K. et al. (2020). "Multicriteria Ammonia Plant Assessment for the Advanced Process Control Implementation". In: *IEEE Access* 8.9260132, pp. 207923–207937. DOI: 10.1109/ACCESS.2020.3038206.
- Ebert-Uphoff, I. and Y. Deng (2012). "Causal discovery for climate research using graphical models". In: *JCLI* 25.17, pp. 5648–5665.
- Ender, D. (1993). "Process control performance: not as good as you think". In: *Control Engineering* 40.10, pp. 180–190.
- Falkowski, M. and P. D. Domanski (2019a). "Residual analysis of nested NARIMA models of the atmospheric distillation column". In: *Proc. of the 18th European Control Conference*, pp. 369–374.
- Falkowski, M. and P. D. Domański (2020). "Impact of outliers on determining relationships between variables in large-scale industrial processes using Transfer Entropy". In: 1, pp. 807–812. DOI: 10.1109/CoDIT49905.2020.9263965.
- Falkowski, M. and P.D. Domanski (2019b). "Nested NARIMA Model of the Atmospheric Distillation Column". In: *AUTOMATION 2019. Advances in Intelligent Systems and Computing* 920.
- Falkowski, Michał J. (2022). "Causality analysis incorporating outliers information". In: *Outliers in Control Engineering. Fractional Calculus Perspective*. Berlin, Boston: De Gruyter, pp. 133–148. ISBN: 9783110729122. DOI: doi:10.1515/9783110729122-007. URL: <https://doi.org/10.1515/9783110729122-007>.
- Falkowski, Michał J. and Paweł D. Domański (2023). "Causality Analysis with Different Probabilistic Distributions Using Transfer Entropy". In: *Applied Sciences* 13.10. ISSN: 2076-3417. DOI: 10.3390/app13105849. URL: <https://www.mdpi.com/2076-3417/13/10/5849>.
- Falkowski, Michał J., Paweł D. Domański, and Ewa Pawłuszewicz (2022). "Causality in Control Systems Based on Data-Driven Oscillation Identification". In: *Applied Sciences* 12.8. ISSN: 2076-3417. DOI: 10.3390/app12083784. URL: <https://www.mdpi.com/2076-3417/12/8/3784>.
- Gao, Z. and S. X. Ding (2007). "Actuator fault robust estimation and fault-tolerant control for a class of nonlinear descriptor systems". In: *Automatica* 43.5, pp. 912–920.
- Gencaga, D., K. H. Knuth, and W. B. Rossow (2015). "A Recipe for the Estimation of Information Flow in a Dynamical System". In: *Entropy*, pp. 438–470.
- Gokas, F. (2000). "Distributed control of systems over communication networks [Ph.D. dissertation]". In: *University of Pennsylvania, Philadelphia, Pa, USA*.
- Granger, C. and P. Newbold (1974). "Spurious regression in econometrics". In: *J Econ* 2, pp. 111–20.
- Granger, C. W. J. (2003). "Time series analysis, cointegration, and applications". In: *Nobel Lecture*, pp. 360–366.

Bibliography

- Granger, CWJ (1969). "Investigating causal relations by econometric models and cross-spectral methods". In: *Comput Math Organ Theory* 37, pp. 424–438.
- Grimble, M. J. (2002). "Controller performance benchmarking and tuning using generalised minimum variance control". In: *Automatica* 38.12, pp. 2111–2119.
- Gupta, M. et al. (2014). *Outlier Detection for Temporal Data*. Morgan & Claypool Publishers. ISBN: 1627053751.
- Hacking, Ian (July 2001). *An Introduction to Probability and Inductive Logic*. ISBN: 9780521772877. DOI: 10.1017/CB09780511801297.
- Hadjiiski, M. and Z. Georgiev (2005). "Benchmarking of Process Control Performance". In: *Problems of Engineering, Cybernetics and Robotics*. Vol. 55. Sofia, Bulgaria: Bulgarian Academy of Sciences, pp. 103–110.
- Hägglund, T. (1999). "Automatic detection of sluggish control loops". In: *Control Engineering Practice* 7.12, pp. 1505–1511.
- Harris, T. J. (1989). "Assessment of closed loop performance". In: *Canadian Journal of Chemical Engineering* 67, pp. 856–861.
- Harris, T. J. and C. T. Seppala (2001). "Recent Developments in Controller Performance Monitoring and Assessment Techniques". In: *Proceedings of the Sixth International Conference on Chemical Process Control*, pp. 199–207.
- Hausman, D. (1998). *Causal Asymmetries*. Cambridge: Cambridge University Press.
- Hawkins, D. M. (1980). *Identification of outliers*. London; New York: Chapman and Hall.
- He, F., C. Wang, and S. K. S. Fan (2018). "Nonlinear fault detection of batch processes based on functional kernel locality preserving projections". In: *Chemometr. Intell. Lab. Lab.* 183, pp. 79–89.
- He, X., Z. Wang, and D. H. Zhou (2009). "Robust fault detection for networked systems with communication delay and data missing". In: *Automatica* 45.11, pp. 2634–2639.
- Heckerman, D., C. Meek, and G. Cooper (2006). "A Bayesian approach to causal discovery. In Innovations in Machine Learning". In: *Springer*, pp. 1–28.
- Hernan, M. A., B. Brumback, and J. M. Robins (2000). "Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men". In: *Epidemiology*, pp. 561–570.
- Hernan, M. A. and J. M. Robins (2020). "Causal Inference". In: *CRC Boca Raton, FL*.
- Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference". In: *J. Comput. Graph Stat.* 20.1, pp. 217–240.
- Horch, A. (1999). "A simple method for detection of stiction in control valves". In: *Control Engineering Practice* 7.10, pp. 1221–1231.
- Hosking, J.R.M. (1990). "L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 52.1, pp. 105–124. ISSN: 00359246. URL: <http://www.jstor.org/stable/2345653> (visited on 09/05/2023).

- Howard, R. and D. Cooper (2010). "A novel pattern-based approach for diagnostic controller performance monitoring". In: *Control Engineering Practice* 18.3, pp. 279–288.
- Imbens, G. W. (2004). "Nonparametric estimation of average treatment effects under exogeneity: A review". In: *Rev. Econ. Stat.* 86.1, pp. 4–29.
- Jafari-Mamaghani, Mehrdad and Joanna Tyrcha (Feb. 2014). "Transfer Entropy Expressions for a Class of Non-Gaussian Distributions". In: *Entropy* 16. DOI: 10.3390/e16031743.
- Jelali, M. (2013). *Control Performance Management in Industrial Automation: Assessment, Diagnosis and Improvement of Control Loop Performance*. London: Springer-Verlag.
- Jiang, H., R. Patwardhan, and S. L. Shah (2009). "Root cause diagnosis of plant-wide oscillations using the concept of adjacency matrix". In: *Journal of Process Control* 19.8. Special Section on Hybrid Systems: Modeling, Simulation and Optimization, pp. 1347–1354.
- Katura, T. et al. (2006). "Quantitative evaluation of interrelations between spontaneous low-frequency oscillations in cerebral hemodynamics and systemic cardiovascular dynamics". In: *Neuroimage* 31.
- Kayser, A. S., F.T. Sun, and M. Desposito (2009). "A comparison of Granger causality and coherency in fMRI-based analysis of the motor system". In: *Hum Brain Mapp* 30.11, pp. 3475–3494.
- Khamseh, S. Afshar et al. (2016). "Control performance assessment based on sensor fusion techniques". In: *Control Engineering Practice* 49, pp. 14–28.
- Knierim-Dietz, N., L. Hanel, and J. Lehner (2012). *Definition and verification of the Control Loop Performance for Different Power Plant Types*. Tech. rep. Institute of Combustion and Power Plant Technology, University of Stuttgart.
- Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger (2004). "Estimating mutual information". In: *Phys. Rev. E* 69 (6), p. 066138. DOI: 10.1103/PhysRevE.69.066138. URL: <https://link.aps.org/doi/10.1103/PhysRevE.69.066138>.
- Kugiumtzis, D. (2009). "Improvement of Symbolic Transfer Entropy". In: *Proceedings of the 3rd International Conference on Complex Systems and Applications*, pp. 338–342.
- LaLonde, R. J. (1986). "Evaluating the econometric evaluations of training programs with experimental data". In: *Am. Econ. Rev.*, pp. 604–620.
- Landman, R. et al. (2014). "Fault propagation analysis of oscillations in control loops using data-driven causality and plant connectivity". In: *Computers and Chemical Engineering* 71, pp. 446–456.
- Lang, X. et al. (2020). "Median ensemble empirical mode decomposition". In: *Signal Processing* 176.
- Laouti, N., N. Sheibat-Othman, and S. Othman (2011). "Support Vector Machines for Fault Detection in Wind Turbines". In: *Proceedings of the 18th World Congress of The International Federation of Automatic Control, Milano, Italy, 28 August-2 September*, pp. 7067–7072.

Bibliography

- Lee, J. et al. (2012). "Transfer entropy estimation and directional coupling change detection in biomedical time series". In: *BioMedical Engineering OnLine* 11.19, pp. 1315–1321.
- Li, J., O. R. Zaiane, and A. Osornio-Vargas (2014). "Discovering statistically significant co-location rules in datasets with extended spatial objects". In: *DaWaK*, pp. 124–135.
- Li, Y. and Z. O'Neill (2015). "Evaluating control performance on building HVAC controllers". In: *International Building Performance Simulation Association*. Hyderabad, India, pp. 962–967.
- Lindner, B., L. Auret, and M. Bauer (2017). "Investigating the Impact of Perturbations in Chemical Processes on Data-Based Causality Analysis. Part 1: Defining Desired Performance of Causality Analysis Techniques". In: *IFAC-PapersOnLine* 50.1. 20th IFAC World Congress, pp. 3269–3274.
- Liu, K., YQ Chen, and P. D. Domański (2020). "Control Performance Assessment of the Disturbance with Fractional Order Dynamics". In: *Nonlinear Dynamics and Control*. Ed. by W. Lacarbonara et al. Cham: Springer International Publishing, pp. 255–264.
- Liu, K. et al. (2018). "A Novel Method for Control Performance Assessment with Fractional Order Signal Processing and Its Application to Semiconductor Manufacturing". In: *Algorithms* 11.7, p. 90.
- Lizier, Joseph T. (2014). "JIDT: An Information-Theoretic Toolkit for Studying the Dynamics of Complex Systems". In: *Frontiers in Robotics and AI* 1. ISSN: 2296-9144. DOI: 10.3389/frobt.2014.00011. URL: <https://www.frontiersin.org/articles/10.3389/frobt.2014.00011>.
- Lynch, C. B. and G. A. Dumont (1996). "Control loop performance monitoring". In: *IEEE Transactions on Control Systems Technology* 4.2, pp. 185–192.
- Mandic, D. P. et al. (2013). "Empirical mode decomposition-based time-frequency analysis of multivariate signals: The power of adaptive data analysis". In: *IEEE Signal Processing Magazine* 30.6, pp. 74–86.
- Mani, S. and G. Cooper (2000). "Causal discovery from medical textual data". In: *Proceedings of the AMIA Symposium* 542.
- Marinazzo, D., M. Pellicoro, and S. Stramaglia (2006). "Nonlinear parametric model for Granger causality of time series". In: *Phys. Rev. E* 73.
- Marschinski, R. and H. Kantz (2002). "Analysing the information flow between financial time series - an improved estimator for transfer entropy". In: *Eur. Phys. J. B* 30, pp. 275–281.
- Mehrotra, K. G., C. K. Mohan, and H. M. Huang (2017). *Anomaly Detection Principles and Algorithms*. Terrorism, Security, and Computation. Springer Publishing Company, Inc.
- Mert Kantar, Yeliz (Nov. 2015). "Generalized Least Squares and Weighted Least Squares Estimation Methods for Distributional Parameters". In: *Revstat - Statistical Journal* 13, pp. 263–282. DOI: 10.57805/revstat.v13i3.176.

- “Method of Moments” (2008). In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, pp. 348–350. ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_262. URL: https://doi.org/10.1007/978-0-387-32833-1_262.
- Miller, David and Martin Kenneth Starr (1969). *Executive decisions and operations research*. 2d ed. Prentice-Hall international series in management. Englewood Cliffs, N.J: Prentice-Hall.
- Mokhov, I. I. and Smirnov D. A. (2006). “El Nino-Southern Oscillation drives North Atlantic Oscillation as revealed with nonlinear techniques from climatic indices”. In: *Geophys. Res. Lett.* 33.
- Nesic, Z. et al. (1997). “CD Control Diagnostics Using a Wavelet Toolbox”. In: *Proceedings CD Symposium, IMEKO*. Vol. XB, pp. 120–125.
- Osborne, J. W. and A. Overbay (2004). “The power of outliers (and why researchers should ALWAYS check for them), Practical Assessment”. In: *Research, and Evaluation* 9, pp. 1–8.
- O’Neill, Z., Y. Li, and K Williams (2017). “HVAC control loop performance assessment: A critical review (1587-RP)”. In: *Science and Technology for the Built Environment* 23.4, pp. 619–636.
- Park, Y. J., S. K. S. Fan, and C. Y. Hsu (2020). “A Review on Fault Detection and Process Diagnostics in Industrial Processes”. In: *Processes* 8.9.
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press, Second ed.
- Pearson, R. K. (2005). “Mining imperfect data; dealing with contamination and incomplete records”. In: *Philadelphia, PA, USA: SIAM*.
- Pikovsky, A., M. Rosenblum, and J. Kurths (2001). “A Universal Concept in Nonlinear Sciences”. In: *Cambridge University Press*.
- Pillay, N. and P. Govender (2014). “A Data Driven Approach to Performance Assessment of PID Controllers for Setpoint Tracking”. In: *Procedia Engineering* 69, pp. 1130–1137.
- (2017). “Multi-Class SVMs for Automatic Performance Classification of Closed Loop Controllers”. In: *Journal of Control Engineering and Applied Informatics* 19.3, pp. 3–12.
- Price, H. (1991). “Agency and probabilistic causality”. In: *British Journal for the Philosophy of Science* 42, pp. 157–176.
- Prichard, D. and J. Theiler (1995). “Generalized redundancies for time series analysis”. In: *Physica D* 84, pp. 476–493.
- Prokopenko, M. and J. T. Lizier (2014). “Transfer Entropy and Transient Limits of Computation”. In: *Sci. Rep.*
- Prosper, Harrison (Apr. 2015). “Practical Statistics for Particle Physicists”. In: DOI: 10.5170/CERN-2014-008.
- Qin, S. J. (2012). “Survey on data-driven industrial process monitoring and diagnosis”. In: *Annu. Rev. Control* 36.9, pp. 220–234.

Bibliography

- Rehman, N. and D. P. Mandic (2009). "Multivariate empirical mode decomposition". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, The Royal Society*.
- Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of Los Angeles Press.
- Robins, J. M., M. A. Hernan, and B. Brumback (2000). "Marginal structural models and causal inference in epidemiology". In:
- Rosner, B. (1983). "Percentage points for a generalized esd many-outlier procedure". In: *Technometrics* 25.2, pp. 165–172.
- Rousseeuw, P. J. and A. M. Leroy (1987). *Robust Regression and Outlier Detection*. New York, NY, USA: John Wiley & Sons, Inc.
- Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and non-randomized studies". In: *Journal of Educational Psychology* 66, pp. 688–701.
- Salsbury, T. I. (2005). "A practical method for assessing the performance of control loops subject to random load changes". In: *Journal of Process Control* 15.4, pp. 393–405.
- (2007). "Continuous-time model identification for closed loop control performance assessment". In: *Control Engineering Practice* 15.1, pp. 109–121.
- Salsbury, T. I. and C. F. Alcala (2015). "Two new normalized EWMA-based indices for control loop performance assessment". In: *2015 American Control Conference (ACC)*, pp. 962–967.
- Schlegel, M., R. Skarda, and M. Cech (2013). "Running discrete Fourier transform and its applications in control loop performance assessment". In: *2013 International Conference on Process Control (PC)*, pp. 113–118.
- Schreiber, T. (2000). "Measuring information transfer". In: *Phys Rev Lett* 85.6, pp. 461–4.
- Schumaker, L. (1981). "Spline Functions: Basic Theory". In: *Cambridge University Press*.
- Shardt, Y. et al. (2012). "Determining the state of a process control system: Current trends and future challenges". In: *The Canadian Journal of Chemical Engineering* 90.2, pp. 217–245.
- Shinskey, F. G. (1990). "How Good are Our Controllers in Absolute Performance and Robustness?" In: *Measurement and Control* 23.4, pp. 114–121.
- Silverman, B. W. (1986). "Density Estimation for Statistics and Data Analysis". In: *Chapman & Hall/CRC: London, UK*, pp. 438–470.
- Spinner, T., B. Srinivasan, and R. Rengaswamy (2014). "Data-based automated diagnosis and iterative retuning of proportional-integral (PI) controllers". In: *Control Engineering Practice* 29, pp. 23–41.
- Taleb, N.N. (2020). "Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications". In: *arXiv:2001.10488, Cornell University Library*.

- Tepljakov, A., E. Petlenkov, and J. Belikov (2012). "A flexible MATLAB tool for optimal fractional-order PID controller design subject to specifications". In: *2012 31st Chinese Control Conference (CCC)*.
- Trewnski, A. (2018). "Regulation of a multidimensional industrial plant with an adaptive feedforward control". Master's Thesis. Warsaw University of Technology.
- Verdes, P. F. (2005). "Assessing causality from multivariate time series". In: *Phys. Rev. E* 72.
- Vincente, R. et al. (2011). "Transfer entropy - a model-free measure of effective connectivity for the neurosciences". In: *J Comput Neurosci* 30, pp. 45–67.
- Visioli, A. (2006). "Method for Proportional-Integral Controller Tuning Assessment". In: *Industrial & Engineering Chemistry Research* 45.8, pp. 2741–2747.
- Wang, H. et al. (2009). "Data Driven Fault Diagnosis and Fault Tolerant Control: Some Advances and Possible New Directions". In: *Acta Autom. Sin.* 35, pp. 739–747.
- Whaley, D. L. (2005). "The interquartile range: Theory and estimation". In: *Master's thesis, Faculty of the Department of Mathematics, East Tennessee State University*.
- Wiener, N. (1956). "The theory of prediction". In: *Beckenbach EF (ed) Modern mathematics for engineers, McGraw-Hill, New York* 8.1, pp. 3269–3274.
- Wilk, M. B. and R. Gnanadesikan (1968). "Probability Plotting Methods for the Analysis of Data". In: *Biometrika* 55.1, pp. 1–17. ISSN: 00063444. URL: <http://www.jstor.org/stable/2334448> (visited on 09/05/2023).
- Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
- (2016). *Causation and manipulability*. The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.)
- Wu, Z. and N. E. Huang (2009). "Ensemble empirical mode decomposition: a noise-assisted data analysis method". In: *Advances in adaptive data analysis* 1.01, pp. 1–41.
- Yang, F. and D. Xiao (2012). "Progress in Root Cause and Fault Propagation Analysis of Large-Scale Industrial Processes". In: *Journal of Control Science and Engineering*.
- Yang, F. et al. (2014). "Capturing Connectivity and Causality in Complex Industrial Processes". In: *Springer Briefs in Applied Sciences and Technology*.
- Yin, S., S. Ding, and H. Luo (2013). "Real-time implementation of fault tolerant control system with performance optimization". In: *IEEE Transactions on Industrial Electronics* 61.5, pp. 2402–2411.
- Yin, S. et al. (2012). "A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process". In: *Journal of Process Control* 22.9, pp. 1567–1581.
- Yin, S. et al. (2013). "Data-driven monitoring for stochastic systems and its application on batch process". In: *International Journal of Systems Science* 44.7, pp. 1366–1376.

Bibliography

- Yin, S. et al. (2014). "A Review on Basic Data-Driven Approaches for Industrial Process Monitoring". In: *IEEE Trans. Ind. Electron.* 61, pp. 6418–6428.
- Yu, Z. and J. Wang (2016). "Performance assessment of static lead-lag feedforward controllers for disturbance rejection in PID control loops". In: *ISA Transactions* 64, pp. 67–76.
- Yuan, H. (2015). "Process Analysis and Performance Assessment for Sheet Forming Processes". PhD thesis. Queen's University in Kingston, Ontario, Canada, pp. 1–214.
- Zhang, J., M. Jiang, and J. Chen (2015). "Minimum entropy-based performance assessment of feedback control loops subjected to non-Gaussian disturbances". In: *Journal of Process Control* 24.11, pp. 1660–1670.
- Zhang, Q. et al. (2019). "Performance Assessment of Cascade Control System with Non-Gaussian Disturbance Based on Minimum Entropy". In: *Symmetry* 11.3, p. 379.
- Zhao, Y.M., W.F. Xie, and X.W. Tu (2012). "Performance-based parameter tuning method of model-driven PID control systems". In: *ISA Transactions* 51.3, pp. 393–399.
- Zheng, B. (2007). "Analysis and auto-tuning of supply air temperature PI control in hot water heating systems". PhD thesis. Dissertation of University of Nebraska.
- Zheng, Y., H. Fang, and H. O. Wang (2006). "Takagi-Sugeno fuzzy model-based fault detection for networked control systems with Markov delays". In: *IEEE Transactions on Systems, Man, and Cybernetics B* 36.4, pp. 924–929.
- Zhong, L. (2003). "Defect distribution model validation and effective process control". In: *Proceedings of the SPIE*. Vol. 5041, pp. 5041–5041–8.
- Zhong, M., S. X. Ding, and E. L. Ding (2010). "Optimal fault detection for linear discrete time-varying systems". In: *Automatica* 46.8, pp. 1395–1400.
- Zhou, J. et al. (2018). "Non-Gaussian Systems Control Performance Assessment Based on Rational Entropy". In: *Entropy* 20.5, p. 331.
- Zhou, Y.F. and F. Wan (2008). "A neural network approach to control performance assessment". In: *International Journal of Intelligent Computing and Cybernetics* 1.4, pp. 617–633.