POLITECNICO DI TORINO

Master's Degree in ICT for Smart Societies

Master Thesis

Detection of collusive behavior in electricity market based on data analysis

Supervisor: Tao Huang

Candidate: Yangshihao Li

Academic Year: 2024/25

Acknowledgement

Writing this thesis has felt like playing super hard game, I wanted to press “Alt+F4”! But in the end I made it .

First and foremost, I would like to thank my supervisor, Professor Huang Tao. He not only guided me academically but also offered unwavering support during times of confusion and anxiety.

I am deeply grateful to my parents, my family, and my girlfriend who has been waiting for me all along. You have always been my strongest support, giving me the courage to move forward without fear.

Thanks also to all the friends and classmates I’ve met along the way. Living in a foreign country, every encounter is a stroke of fate.

Finally, thank you to my little laptop. Even without a GPU, you've always run like a champ and never let me down.

Author:

Yangshihao Li

Abstract

The electricity system is very important for daily life. People can not live without electrical devices. Society and entire countries depend on electricity too. It can help the economy grow, make society better, and support new technologies. Today's electricity market has changed a lot from before. Many countries now pay attention to protecting the environment and are developing clean energy sources like wind, water, and solar energy. These big changes affect the electricity market first. Keeping the market healthy is important for social stability and protecting people's money. That's why we need to control harmful actions in the electricity market.

We chose Italy's Day-Ahead Electricity Market (MGP) for our study. We want to use statistics, machine learning, and economic ideas to give early warnings about harmful actions in the market. Chapter One gives background information about how electricity markets started, how they changed, and how Italy's electricity market looks now. It also talks about the challenges and future directions of the electricity market. Chapter Two covers basic electricity market knowledge and explains how to recognize market power and collusion, using real cases to show the serious effects of collusion.

Chapter Three introduces the methods and ideas we used, explaining why we chose them. Chapter Four explains each step in detail, from getting original data to calculating indicators of market power, giving a complete analysis of Italy's electricity market. Based on the special characteristics of collusion, we chose suitable methods to find unusual activities.

Chapter Five looks at the results from our model and combines them with extra information. Although some unusual results can't be fully explained, there is not enough evidence to say they are definitely collusion. But the model helps narrow down possible collusion cases and gives early warnings. Finally, Chapter Six summarizes our findings and offers suggestions for future research.

Keywords: Electricity market, Collusive behavior, Data analysis, Anomaly detection

Table of Contents

Table of Figures

Table of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 The Development of Electricity and the Power Market

In 1882, Edison built the Pearl Street Power Station in Manhattan, New York. People see this as the world's first successful commercial electricity system. This also marked the start of electricity moving from experiments in labs to large-scale use. Over the next few decades, cities grew quickly, and the demand for electricity in industries also rose steadily. Because of this, electricity use in the United States grew very fast, with an average yearly growth rate of 25.7% between 1892 and 1910, according to Nye's research in 1990[1]. Such rapid growth led to a market structure based on vertical integration. By 1935, three large holding companies controlled up to 83% of the country's electricity. This shows that the industry was very concentrated at that time, as studied by Hughes in 1983.

At that time, this high centralized system can helped build electric grids quickly and reduced regional electricity shortages. But, people started noticing problems caused by monopolies, such as "X-inefficiency" . Joskow's [2] research in 1974 clearly tell us that monopoly electricity companies had costs about 18% to 27% higher than those in competitive markets. This finding was different to the traditional idea. They thought that bigger companies were always more efficient. Joskow provided the theoretical basis for later reforms aimed at making electricity markets more competitive.

1.1.2 Market Liberalization in the Electricity Sector

At first, electricity market reforms were expected to open up the industry, break monopolies, and improve the efficiency of resource allocation. Policymakers believed that bringing in market competition could help lower electricity prices, drive innovation, benefit consumers, and support overall economic growth. But there is often a big difference between ideas and reality. With the opening up of electricity market, some new problems came out. Such as price manipulation due to too much market power, unstable electricity supplies because of generation capacity changing, and market chaos caused by lack of regulation[3].

In this situation, the regulators become more and more important. They don’t just only make the rules anymore and leave them alone. They also need to know how the market works and try to stop problems before they grow. It is very hard to find a balance between encouraging competition and stopping unfair behavior. In fact, opening up electricity market hasn’t been a easy process.It’s more like a long experiment, need to adjust rules and whole market all the time.

1.1.3 The Economic Impact of the Electricity Market

Electricity is not just important for the economy. It also plays a key role in government planning, public services, and improving industries.

This is especially true in developing countries. These places are often going through big changes in their economy or have weak infrastructure. In such cases, electricity becomes even more important. A good example comes from India. A study by Burlig and Preonas in 2016 looked at the effects of bringing electricity to rural areas. They found that villages with electricity had 1.8 times more agricultural output compared to those without it[4]. This shows how better electricity access can help raise productivity and support development.

But electricity influence is much bigger than economic output. Today, it is very important to build a stable and reliable electricity system. It helps factories and companies keep running smoothly and supports the whole supply chain.It lets prices guide resources.This can attract foreign investment and make a country more competitive.On top of that, having a steady power supply is linked to energy security. And energy security is an important part of keeping the whole economy stable in the long run.

At the same time, changes in electricity markets are closely connected to global environmental issues.As renewable energy becomes more common, old systems that depend on fossil fuels are slowly being replaced. New systems are greener and smarter.This shift helps reduce emissions and fight climate change.It also pushes the development of new technologies and industries. Examples include energy storage, smart power grids, and carbon trading systems.

From this point of view, building a healthy electricity market is not just about being efficient.It also related with protecting the environment, making sure the system is fair, and meeting national goals.In order to achieve this, it is important to study how the market works. So that we can find out when it comes to unfair actions like manipulation or collusion.These problems can make people do not trust in the market. Stopping them helps keep the system fair.

1.2 Italian Electricity Market

1.2.1 The Development of the Italian Electricity Market

The way Italy's electricity market has changed shows bigger changes in its energy policy, market setup, and views on the environment.. As shown in Figure 1-1, the market started in the late 19th century.After that, it went through a period of strong national control.In recent years, it has moved toward more open competition and greener energy.Overall, the development of Italy's electricity market has been complex, but also shows key trends that many other countries may face too.

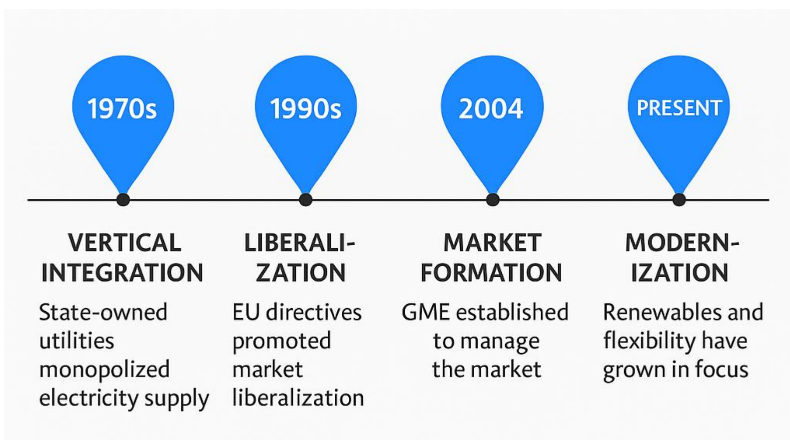

Figure 1-1 Italian electricity market development

1.2.1.1 Early Development of Electricity(1800s-1960s)

In the beginning, most of Italy's electricity systems were run by local private companies.These companies mainly provided power for city lights and tram lines[5]. After World War II, during the rebuilding period, the government started to see electricity as very important.Because of this, the state began to take more control.Over time, new policies pushed the system toward central management.

1.2.1.2 Nationalization and the Formation of ENEL(1960s-1990s)

In 1962, Italy set up ENEL (Ente Nazionale per l'Energia Elettrica).This marked the start of a new period, where the national electricity system was run in a more unified and centralized way[6]. This model helped bring electricity to more people and expanded the power grid across the country.But over time, problems started to show up.The system became less efficient, and the market lacked flexibility and competition.

1.2.1.3 Liberalization of the Electricity Market(1990s-2000s)

In the 1990s, the European Union asked their member countries to create a energy market together. Italy started to open up its electricity sector for response.This was known as "Bersani Decree."[7]. The new rules allowed private companies to take part in power generation and sales.That was the reason that ENEL's monopoly was broken down.At the same time, customers had more choices.This rule pushed companies to improve their services and let market forces shape electricity prices.

1.2.1.4 Promotion of Renewable Energy and Modernization of the Electricity System(2000s till now)

In the 21st century, many countries agreed that the energy system needed to change. This pushed Italy to speed up the use of renewable energy. Wind and solar power received strong support from the government. One example is the "Conto Energia" program[8] , which helped grow the solar power market a lot. At the same time, Italy started using smart monitors more widely. Making the market more digital can also improved the electricity grid. These changes made the system smarter and more open. Now days, Italy's electricity market is still improving. New technologies and policies can protect this market very well.

1.2.2 Italian Electricity Market Structure

Figure 1-2 shows the structure of Italy's electricity market. The market is made up of several smaller parts. These parts have different roles but work closely together. Together, they form a complete and well-organized system. The most important parts are the day-ahead market (MGP), the intra-day market (MI), and the ancillary services market (MSD). Each one is key to making sure electricity is scheduled, balanced in real time, and the whole system stays stable.

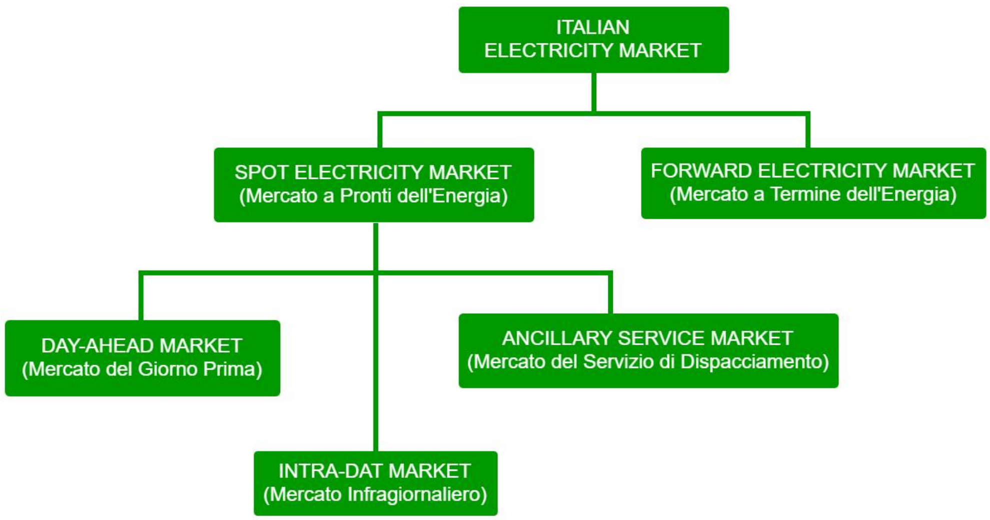

Figure 1-2 Italian electricity market structure

The day-ahead market is the largest and most structured part. It is where most of the electricity planning happens. Producers and consumers submit their bids one day before the electricity is actually delivered. These bids are based on forecasts of demand and prices. The market clears through an auction. A single market price is set where supply meets demand. This system helps everyone plan better and gives a clear price signal for running the grid.

Compared to the day-ahead market, the intra-day market focuses more on flexibility and quick response. It helps fix the gaps caused by forecasting errors from the day before. On the day of delivery, producers and consumers can adjust their plans. Because everything can be changed in real time, such as power generation, changes in demand, or technical problems. This market is very useful for a lot of renewable energy. That also because wind and solar power can change quickly, this market helps us to manage those short-term changes.

The ancillary services market is different but also very important . It helps us to keep the system stable. This market manages frequency control, voltage support, and backup power (reserve capacity). These services don't directly affect electricity prices, but they are key to keeping the grid running safely. In Italy, these tasks are mainly handled by Terna, the national transmission system operator.

It's important to note that Italy's electricity market is divided into seven main zones. These zones are based on differences in supply and demand across regions, as shown in Figure 1-3[9]. The northern and central regions have large populations and many industries. As a result, they use a big part of the country's electricity. On the other hand, the southern regions and islands rely more on renewable energy or local power generation. This regional setup helps make electricity prices more accurate. It also allows the system to respond better to local imbalances between supply and demand.

Figure 1-3 Map of market zone (source: [9])

1.3 Current Challenges and Future Trends

1.3.1 Key Challenges in the Current Market

As more countries open up their electricity markets and move toward cleaner energy, new challenges are starting to appear. One big issue is the growing difficulty in keeping supply and demand balanced. This problem is worse when the system depends heavily on wind and solar power.

That makes electricity output less stable because these kind of energy change with the weather. Technology bring energy systems more efficient, and also brings new risks. For example, many people are concerns about data privacy and cybersecurity. These risks are even bigger in systems for now days because of using digital tools.

Although liberalization can make markets more efficient, but also bring problems. Without government supervision, big companies might use their power to increase prices for no reason. This is unfair competition and hurting markets.

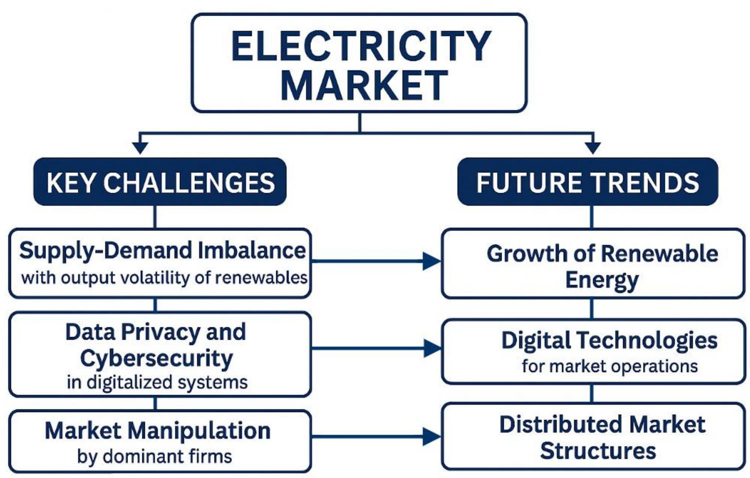

Figure 1-4 Key challenges and future trends

1.3.2 Technological Innovation and Market Adaptation

A new generation of technologies is fundamentally reshaping the operational logic of electricity markets. The advancement of smart grid infrastructure has significantly improved the responsiveness of dispatch systems, while also increasing the transparency and controllability of energy flows[10]. At the same time, better energy storage systems are helping to face with challenges of renewable energy.

These changes in technology are also pushing the market to evolve. Every part of the system needs to improve. This includes from the generation side to the end users. The system must be more efficient. It also needs to be safer and easier to adjust . Whether the market and regulators can keep up with these changes is very important. It will determine how stable and reliable the electricity system will be in the future.

1.3.3 Future Trends

Looking ahead, electricity markets will focus more on being sustainable and smart. On one hand, as the world keeps working to fight climate change, the use of renewable energy will keep rising. It is expected to become one of the main energy sources. On the other hand, digital tools like automated dispatching, block-chain, and virtual power plants will become more important. These technologies will help run the market more smoothly.

At the same time, new policies will support a shift from centralized systems to more flexible and local ones. In the future, electricity markets will likely be more connected with technology, better regulated, and more coordinated overall.

1.4 Research Motivation and Objectives

As the markets are opening up and turning greener, they are also becoming more complex. Companies in the market not only just respond to prices. They may also use some strategies or even work together in secret. By using information difference, this kind of behavior becomes harder to find. Some companies may try to influence prices by changing how they supply electricity or by placing bids together. If this behavior goes unnoticed, it can cause hurt fairness, lower efficiency, and make the system less stable.

Many past studies have focused on theories or specific case examples. But there are still not many studies that use large amounts of real trading data and modern algorithms. We arr going to focus at the Italian day-ahead electricity market, and uses two methods:

frequent pattern mining and sequence modeling. The goal is to find patterns of trading behavior that happen very often in the same way. These patterns might point to possible collusion.

This study does not only try to judge market behavior. Instead, it builds a tool that can help market regulators and analysts. By using data to find and show hidden patterns, this research aims to support future work on finding collusion. It also hopes to give new insights into how market players act at a detailed level.

Chapter 2

Market power and collusion in electricity market

2.1 Market power

2.1.1 Definition of Market Power

In electricity markets, market power means the ability of a company to push prices away from fair, competitive levels on purpose. This idea was first explained by economist Lerner in 1934[11]. He said that when a company can raise prices without losing customers, it shows it has market power. Electricity markets are more complicated than regular commodity markets. This is because of the special way electricity works and how the system is run.

First, electricity always used in real time, so the supply and demand must match at every second. But electricity is hard to store, and users usually don't change their using habits quickly when prices change. This makes prices unstable when supply suddenly changes. Economist Wolak studied this issue in California's electricity market[12]. He showed that a few generators were able to take advantage of these sudden shifts to raise prices. Second, the way electricity transmitted across the grid can also cause local problems. Some areas have transmission limits, only one or two generators can send power. This means they can act like a monopoly, even if there are still many players in market. Borenstein and colleagues (2000) explained how these limits give local units more power to raise prices without competition[13]. Third, power stations can not change the productions quickly. Some may have a minimum amount that they must generate, which limits their flexibility. This problem cannot be handled, because companies can't easily respond by changing output. Cramton and Stoft (2005) discussed this when they argued for smarter capacity markets that account for physical limits[14].

A 2021 report by the Agency for the Cooperation of Energy Regulators (ACER) found something surprising. Even companies with less than 20% market share were able to raise prices by over 35% in areas with transmission congestion[15]. Even companies with less than 20% market share were able to raise prices by over 35% in areas with transmission congestion. This shows that market power isn't only about how big a company is. It also depends on where they are, how the grid works, and how they behave. Because of this, traditional indicators like CR4 or HHI, only measure market size. They are not good enough for full conditions. So we need better tools which can consider both time and location conditions, and that can track behavior when it changes.

2.1.2 Market Power Behaviors

In theory, market power means a company can keep prices above marginal cost for a long time. But in real life, market power is often seen through how companies behave. There are three common ways this happens: raising prices, hiding capacity, and blocking new competitors.

First, pricing is the most obvious way to use market power. Some big companies may raise their bids during times of high demand, such as very cold winters or hot summers. They do this to take advantage of the tight supply. This allows them to earn extra profit. Even if this behavior follows market rules on the surface, it often goes against the idea of fair competition.

Second, controlling capacity is a common way companies use market power. Big power producers might hold back part of their capacity on purpose. They may also delay starting their plants. This makes it look like there is not enough supply. As a result, market prices go up. In some cases, companies raise their bids too high on purpose. This forces more expensive units to be used earlier, which helps them earn more profit. These actions may seem reasonable from a business point of view, but their real purpose and effect deserve careful attention.

Finally, blocking new players is a more hidden way of using market power. Some companies sign a long time contracts that don't let others enter the market. Others will raise the cost of connecting into the grid. Or they require new players to pay for expensive infrastructure. These strategies don't change prices right away. But they make it harder for new competitors to join. In the long-term, this hurts market openness and reduces competition.

2.1.3 Market Power Indicators

Quantifying market power is a fundamental and essential task in electricity market regulation. In the following section, we introduce several commonly used indicators and explain their calculation methods.

Herfindahl-Hirschman Index (HHI)

$$HHI = \sum_{i=1}^{N} (s_i)^2 \times 10,000 \tag{2.1}$$

In formula (2.1) , s_i denotes the market share of firm i, expressed as a decimal (e.g., 0.3 for 30%). N is the total number of companies in the market.

≤ 1000	Low Risk
$1000 \sim 1800$	Median Risk
> 1800	High Risk

Table 2-1 Score ranges of HHI

The Herfindahl-Hirschman Index (HHI) is a simple but effective way to describe how concentrated a market is. In Table 2-1, it tells us whether a few companies dominate the market or if competition is more evenly spread out.

Concentration ratio

The Concentration Ratio (CR) is another widely used metric to evaluate how concentrated a market is. Unlike HHI, which takes into account all companies and squares their market shares, CR focuses only on the top n companies—for example, CR4 looks at the combined market share of the four largest companies.

$$CR_n = \sum_{i=1}^{n} s_i \tag{2.2}$$

<14%	Low Risk
14% ~ 71%	Median Risk
>70%	High Risk

Table 2-2 Score ranges of CR4

This indicator gives a quick sense of whether a few companies control the market. The higher the concentration ratio, the less competitive the market is likely to be. The perfect CR_4 score from Table 2-2 is less than 14%.

Entropy coefficient

The **entropy coefficient** is a way to measure how evenly market shares are spread among companies. It was introduced by **Hart** in 1971[16].

$$E = -\sum_{i=1}^{N} s_i \cdot \log(s_i) \tag{2.3}$$

Where s_i is market share of company i, N is number of companies.

If all companies have similar market shares, the entropy value will be high, indicating a competitive and well-balanced market. If a few companies control most of the market, the entropy value will be low, suggesting higher concentration.

Entropy is especially useful when the market has lots of small companies. HHI and CR might miss their presence, but entropy picks up on that distribution.

Lerner index

The Lerner Index is a common way to measure a single company's market power. It shows how much a company can raise its price above the marginal cost. In simple terms, the more a company can mark up its price without losing customers, the more power it has in the market.

$$L = \frac{P - MC}{P} \tag{2.4}$$

In formula (2.4), P is the price of the product, and MC is the marginal cost.

Unlike indicators that look at the whole market, the Lerner Index focuses on single company. A higher index value means the company has more control over prices. This also means it holds a stronger position in the market.

2.2 Collusive Behavior

2.2.1 Definition of Collusive Behavior

In many countries and regions, collusion is seen as a serious violation of fair competition rules. This kind of behavior can lead to price manipulation, blocking other players from the market, and poor use of resources. In the end, it hurts consumers and reduces market efficiency.

In electricity markets, things are more complex. These markets are sensitive and have many technical rules. Because of this, collusion is often harder to spot. As Stigler (1964)[17] explained, collusion happens when two or more companies work together in secret. They do this to change prices or output levels on purpose so that they can get extra profit. Usually, there are no contacts. Instead, they use quiet signals or copy other's behavior. It looks like they are making decisions on their own. But, their actions are coordinated.

In the Treaty on the Functioning of the European Union (TFEU)[18], Article 101 clearly defines what counts as collusion.

There are three main points in this legal definition:

Collusive intent: This includes clear agreements, whether spoken or written, or silent understandings based on behavior, signals, or information sharing.

Purpose or effect of restricting competition: Even if the goal was not to reduce competition, any actions that leads to less competition can be seen as collusion..

Actual market harm: This often shows up as strange price changes, lower efficiency, or long-term market problems.

The key feature of collusion is that it relies on hidden cooperation and non-transparent information sharing. It is difficult to detect than open price-fixing. So regulators need more than just structural indicators like market share. They must also pay more attention to patterns in behavior, and the market also changes over time.

2.2.2 The Performance of Collusive in Electricity Markets

Because electricity can't be stored easily, and demand doesn't change much with price, the market can change quickly. These features make collusion in electricity markets more hidden and more strategic. Based on past studies, such as Fabra (2003) [19], we can divide collusion into three main types.

(1) Explicit collusion: This collusion is the easiest to find. In this case, companies make secret deals with each other. For example, they may agree to raise prices together at some times. Or they may help each other during peak hours to push prices higher. In some situations, companies may lie about grid congestion condition. This makes it look like there is not enough capacity, which push up prices. These actions help them earn extra profits by creating a fake news.

(2) Tacit collusion: This type does not need direct communication between companies. They observe other's behavior and adjust their own actions. For instance, a large company might give a very high bid to test the market. Other small companies may follow with similar bids. Even though each company seems like acting alone, but in the end their prices are very close. This is silent coordination. Because there's no clear agreement, this kind of collusion is much harder for regulators to detect.

(3) Structural collusion: This happens with only a few companies and a stable structure. In these cases, companies may quietly divide the market by themselves. For example, they may focus on different regions, customer types, or load levels. This reduces direct competition. It also allows each company to control its own part of the market, like a local monopoly.

2.2.3 Harms of Collusive Behavior

Collusion not only harm the companies involved, but also hurts consumers and the all market.

The first and most direct result, the electricity becomes more expensive. In a normal market, prices should be set by supply and demand. But when a few companies work together, they can control prices. This leads a higher price to buy electricity. For families and small businesses, that means bigger bills and higher running costs.

Second, collusion makes the supply system less efficient. As mentioned before, some companies may reduce their output on purpose.It will let customers consider that there's not enough electricity. This kind of behavior wastes resources. It can also cause power shortages during times of high demand.

Third, collusion influence innovation. In a fair market, companies try to improve themselves to stay ahead. They invest in better technology and try to be more efficient. But when the market is quietly divided, they don't need to compete. They focus on keeping profits without improving. That slows down progress in the industry over the time.

On a bigger scale, collusion damages fairness and openness in the market. It leads to poor use of resources. People lose trust in the system. It also makes it harder for policies to work as planned. In the long term, this brings large costs for everyone.

2.3 Collusion in the Electricity Market

2.3.1 Case One: The California Electricity Crisis (2000–2001)

Between 2000 and 2001, there was a serious electricity crisis in California. Several power companies used dishonest methods to manipulate the market. For example, they reported station maintenance on purpose. This made it seem like lack of electricity. They also pretended to make power trades between states, even there was no electricity moved. These tricks created shortage. As result, prices went up very quickly. During the worst times, electricity prices jumped from around \$30/MWh to over \$300/MWh. That's a ten times increase. This huge price increase caused a major budget shortage in the state. According to a **California Senate report from 2002**[20].

Many households and small businesses were hit hard. Their electricity bills went up sharply. In some areas, there were rolling blackouts, where power was cut off on purpose. Some small businesses had to shut down. This crisis shook public trust. People lost confidence in the electricity market and also in the government's ability.

This case also shows that manipulation is not limited to one area. In this case, companies used cross-regional trades to trick the system. This made the problem even bigger, with effects that went beyond California's borders.

2.3.2 Case Two: Market Allocation and Information Exchange in the EU (2019)

In 2019, the European Union began checking on some big electricity companies. They found that companies like RWE and EDF had made quiet deals. Each company stayed in its own area and didn't try to get projects in the other's country. They didn't compete. Instead, they worked together to keep their part of the market.

These companies also shared price ideas in a more hidden way. They used groups like industry associations to talk about what they expected prices to be. This affected how the market acted. The European Commission said this kind of behavior cut down cross-border electricity trade by at least 10%. It also caused high prices in some countries for a long time[21].

This case shows that collusion isn't just about raising prices. It also includes quiet deals, like splitting the market or sharing price info. These things may seem small. But they still hurt fair competition. And they clearly break rules like Article 101 of the EU treaty.

2.4 Approach to Collusion Detection in Electricity Markets

This part builds on the earlier discussion about market power. It looks at what collusion means, how it happens, and some real cases from electricity markets. As we've seen, collusion is often hard to notice. It can show up as price changes, holding capacity, splitting the market, or sharing information. Sometimes it is done through clear agreements. Other times, it just happens quietly without any deal. Because of this, normal rules can not handle it.

Even though laws like Article 101 of the EU treaty clearly define collusion, but it is still hard to avoid. The electricity data is complex, their behavior is always hidden, and proof is weak. New ways of trading and digital platforms also make things even harder.

To deal with this, more researchers are turning to market data and machine learning. These tools offer better ways to find collusion. In this study, we look at the Italian day-ahead electricity market (MGP). We use data analysis, pattern mining, and machine learning to find early signs of collusion. Then, we check if the results make sense from an economic point of view and fit the real market. The flow chart is as follow Figure 2-1.

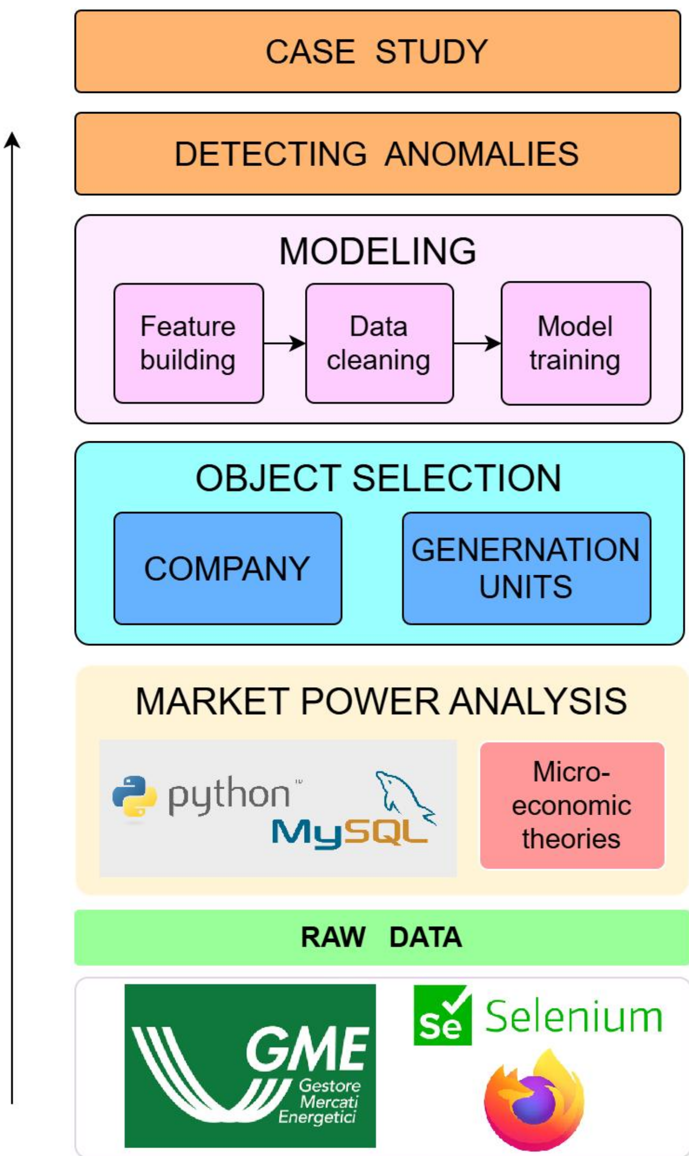

Figure 2-1 Flow chart

Chapter 3

Research Methods and Tools

3.1 Theoretical Basis

3.1.1 Basic Principles of Machine Learning

Machine learning helps computers get better by learning from old data. People do not need to preset rules. It works in large numbers of data.

From Mitchell's research, there are three types of machine learning[22].

Supervised Learning: The input of this part are data with label. After learning their features, models can classify new input by itself.

Unsupervised Learning: This approach deals with data that without labels. Its will find inner relationships within the data, such as clustering, grouping, or discovering hidden features and connections.

Reinforcement Learning: In this part, the learning can be seen as a "game." The system keeps trying everything and watching what happens. It will change their strategies based on the results

3.1.2 Statistical Methods

Statistical methods is very important in this study. They not only help in understanding the hidden features of the electricity market but also provide preprocessing data before model training. After all, it is essential to find relative features before building a model.

For example, we use Pearson correlation coefficient[23] to calculate the similarity between different generating units. It provides a linear indicator between two variables, and the formula is as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3.1}$$

As shown in formula (3.1), this actually calculates the covariance between two variables X and Y. The value is between -1 and 1, the value is bigger, the more similarity.

And we also apply some basic statistical indicators, such as mean, variance, minimum, and maximum. Although these indicators are simple, but they are important for analysis.

3.1.3 Application Areas and Role in Research

Electricity market data are always high dimension, time sequence, and non-linear relationships. These characteristics make it difficult to learn for traditional models. In this situation, we can use statistical methods to help describe and understand the data. After this, we use machine learning to study "normal behavior". Then it can find potential anomalies.

In other words, statistical methods focus on understanding the data. Machine learning focuses on learning from the data. Together, they can give us a complete and complementary analysis pipeline.

3.2 Selection and application of Machine learning tools

3.2.1 Overview of the Python Environment

We need to choose a very strong tool as programming language, when we face a complex data-set. We selected Python because it offers a good balance between flexibility and performance, and provides many useful libraries. It can be well-suited for both data cleaning and model development.

We mainly use python in data processing, modeling, and visualization, the following three Python libraries were used.

Pandas: Very useful in data organization, cleaning, and feature construction.

Matplotlib and Seaborn: Used to create visualizations of curvesand value distributions, helping us to know data in different way.

Scikit-learn[24]: A key library for machine learning mission such as classification, clustering, and dimension reduction.

3.2.2 Common Machine Learning and Modeling Libraries

In this study, we used both Scikit-learn and PyTorch.

Scikit-learn is a general machine learning library in Python, it has many standard algorithms inside. It's suited for quick model testing. PyTorch provides dynamic computation graphs and GPU. It can work with deep neural networks and dealing complex, large-scale data.

Table 3-1 is a brief comparison of the two libraries in terms of user experience and core features:

Feature	Main use	Strengths	Limitations
Scikit-learn	raditional machine learning; fast prototyping	Easy to use, well-documented, active community	No GPU support, not suitable for deep learning
PyTorch	Deep learning, sequence modeling, complex structures	High flexibility, GPU support, customizable layers	More setup needed, steeper learning curve

Table 3-1 Comparison of Scikit-learn and PyTorch

3.3 Data Management

3.3.1 MySQL Database

Because the electricity data is large and complex, we chose MySQL to store it. MySQL is a free database that many people use. It is stable, can grow with more data, and works well when there are many queries. It is also good at keeping data safe and correct[25].

In our project, MySQL helped us make flexible tables and use indexes to speed up data search. It stored big data quickly and kept the data correct. MySQL was also easy to use, which matched the needs of our work.

3.3.2 Application in Data Processing

Throughout the data processing workflow, MySQL was used to centrally manage a large volume of structured transaction data collected from the Italian MGP market. Key fields included transaction time, bid prices, volumes, unit IDs, and market participants.

We first cleaned and standardized the raw data, then imported it into the database to ensure structural consistency for later analysis. While working with the data, we used SQL to get the parts we needed. For example, we could pick data by time, price, or user ID. This gave us the right data to train our model [26].

When we added more data, the database will update and store the data step by step.

.

3.3.3 Data Security and Backup

Data security is important part during the analysis process. Some data in electricity market maybe are private, so we can not lose the data by mistake, we set up MySQL's permission system carefully. And must use password to access database everytime.

We also use automatic backup with the mysqldump tool. Then generate data snapshots periodically. After that we can quickly find back data from system failures or errors[27].

We also turned on encrypted connections when log in MySQL. These steps made the data system safe and more stable.

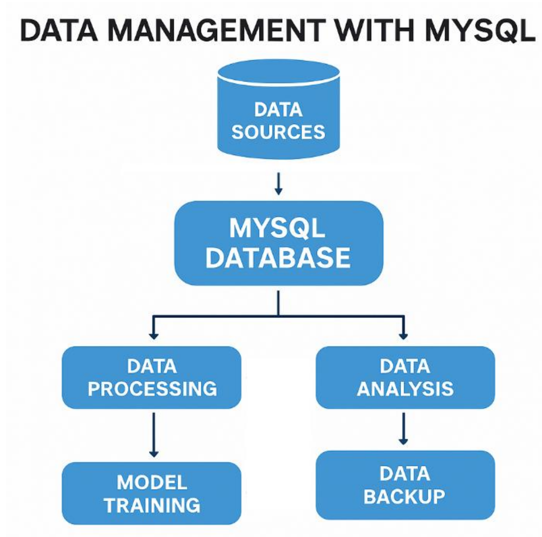

Figure 3-1 MySQL data management

3.4 Algorithm and Model Design

To detect potential collusion in electricity markets, we combined two different principles. The Eclat algorithm for rule mining. Transformer-based auto-encoder (Transformer-AE) for anomaly detection.

Eclat can find the most frequency behavior pattern. It will find information in large number of data. It can help us points to coordinated actions. That actions may be collusive between units.

The Transformer-AE focuses on modeling, learn normal behaviors and find abnormal behaviors deviations. It is an unsupervised learning model, it does not rely on label data. Transformer_AE provides a more dynamic and detailed trading activity.

3.4.1 Association Rule Mining (Eclat)

Eclat is an old method used to find common item groups. It was made by Zaki in 1997[28]. This method will use list to store whether data is appear. Then compute these appearing sets, we can get a support score for each set. This can help us find patterns in large dimensional data.

Eclat is similar to Apriori algorithm, but Eclat can remember more information. So it has a good performance in large-scale structure data-set. In our study, we use Eclat to find these generator units which have similar actions. For example, if two units always bids at similar prices in the same time slots, we can consider this might be coordination.

Algorithm 1 Eclat Algorithm

Require: Transaction ID sets T, minimum support $minsup$
Ensure: Frequent itemsets

for each item i in T **do**
  if support$(i) \geq minsup$ **then**
    Report $\{i\}$
    Eclat$\{i\}$, $T[i]$
  end if
end for
Eclat$prefix$, $tidlist$
for each item j after $prefix$ **do**
  $new \leftarrow tidlist \cap T[j]$
  if support$(new) \geq minsup$ **then**
    $X \leftarrow prefix \cup \{j\}$
    Report X
    Eclat$X$, new
  end if
end for

Algorithm 1

Algorithm 1 shows that how Eclat works. We filter out this similar actions with a minimum support threshold. we filter out frequent co-occurrence patterns as early warning signals of possible collusion. This result will be part of input feature for Transformer-AE.

3.4.2 Transformer Auto-Encoder Model

Auto-Encoder and Anomaly Detection Mechanism

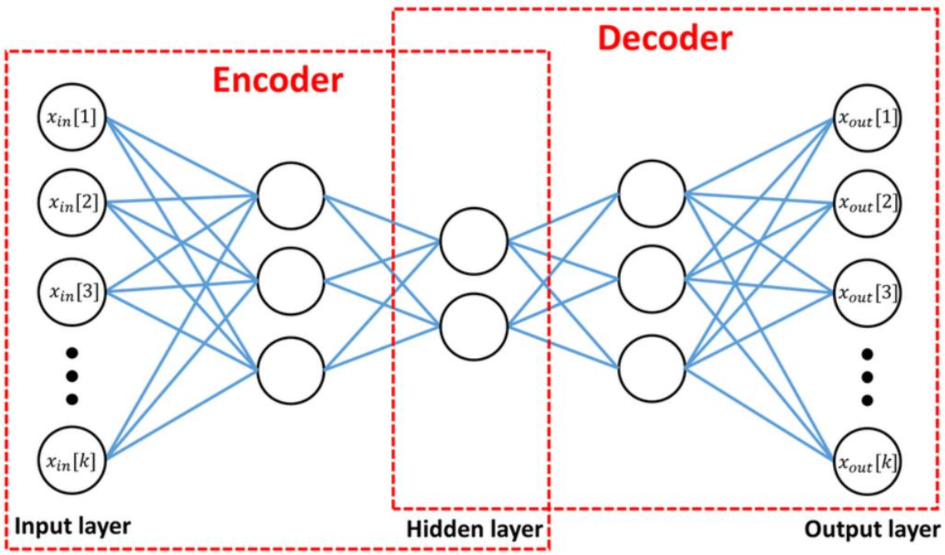

Figure 3-2 Auto-encoder structure (source: [29])

An auto-encoder is a type of model that can be trained without using labeled data. It's usually made up of two parts in Figure 3-2[29]. An encoder that compresses the input, and a decoder that tries to rebuild the original input from that compressed form. The main idea is to learn and capture the structure of the data by shrinking it and then putting it back together as closely as possible.

Once the model is trained, if it struggles to rebuild a new input and the error is much higher than the average during training, it's a sign that something might be off. Algorithm 2 is a pseudo-code of anomaly detection. In this study, these cases may point to possible market manipulation or collusion.

Algorithm 2 Autoencoder-based Anomaly Detection

Require: Normal dataset X, Anomalous dataset $x^{(i)}$ for $i = 1, \ldots, N$, threshold α
Ensure: Reconstruction error $\|x - \hat{x}\|$
Train an autoencoder using the normal dataset X
Let ϕ be the encoder, θ be the decoder
for $i = 1$ to N **do**
    error$(i) \leftarrow \left\|x^{(i)} - g_\theta\left(f_\phi\left(x^{(i)}\right)\right)\right\|$
    if error$(i) > \alpha$ **then**
        $x^{(i)}$ is an anomaly
    else
        $x^{(i)}$ is not an anomaly
    end if
end for

Algorithm 2

Transformer self Attention Mechanism

The self-attention mechanism in Transformer is a basic and important method. It helps the model catch more detail information, and judge which ware more important. In other words, the model can choose specific feature pay more attention to. Figure 3-3[30] shows the structure of self attention mechanism.

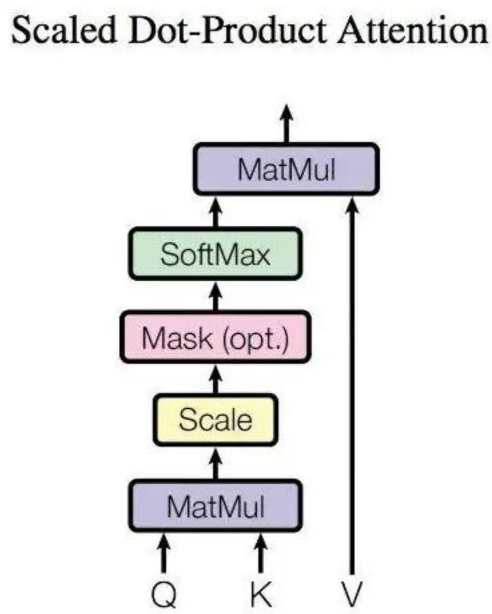

Figure 3-3 Self attention structure (source: [30])

We can use linear computation between input data and weight. The matrices W in the formulas (3.2) represent weight parameters. Then we can get three vectors Q,K,V, the model will set this weight itself.

$$Q = XW^{Q}, \quad K = XW^{K}, \quad V = XW^{V} \tag{3.2}$$

Here is the formula for self-attention.

$$Attention(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{3.3}$$

Multi-head attention is an other form of self attention. It can be seen as multiple self-attention mechanisms together. And each head calculate their score independently. In that way model can consider more information from different features and relations. Figure 3-4 [30] shows the structure of Multi-head attention.

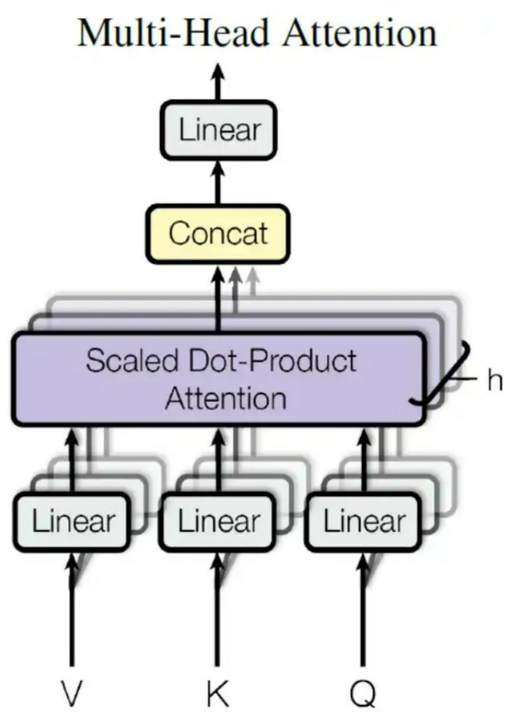

Figure 3-4 Multi-head attention structure (source: [30])

(3.4) shows the multi-head attention formula

$$MultiHead(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{3.4}$$

Structure of Transformer Auto-Encoder Model

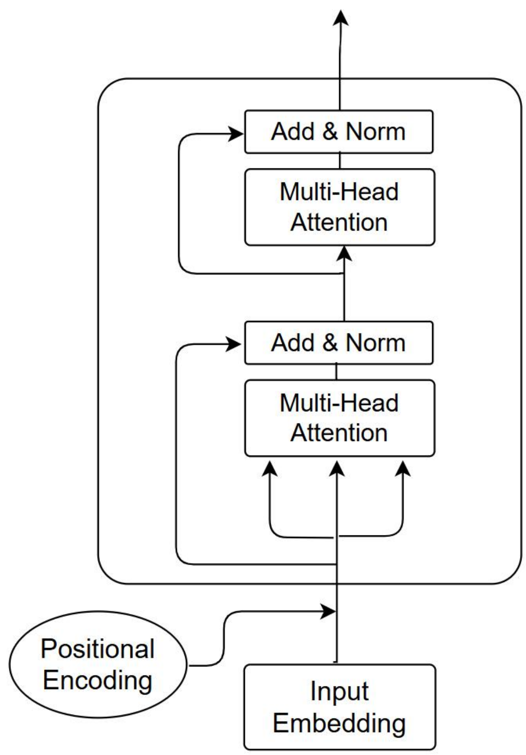

Figure 3-5 Transformer encoder structure

The Transformer Encoder shown in Figure 3-5. It is made of several layers. Embedding layer is the first layer. It changes input data into higher-dimensional vectors so that model can understand easily. Next is the Positional Encoding layer. It used to give sequence number to input data. Because Transformer model can not handle positioning data.

After that is the Add & Norm, it add residual information (Add) and normalize data (Norm). Seen from Figure 3-6.

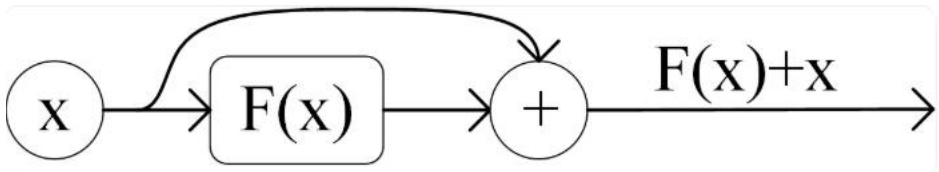

Figure 3-6 Principle of Add&Norm layer

This is for let the training process stable. So there will not occur vanishing gradients. Following is the Feed Forward network. It includes two linear layers and nonlinear activation function (such as ReLU) .

Finally, we used a MLP as the Decoder shown in Figure 3 - 8. It is made of several linear layers and activation functions. It receives the features output from the Transformer Encoder, and rebuild them back to their original dimensions. This part can be seen as a small auto-encoder.

The first linear layer increases the feature dimension to add more non-linear information. Then comes the GELU activation function. As we can see from the Figure 3-7[31], compared to ReLU, GELU still gives non-zero values for negative inputs, which helps the model capture subtle changes.

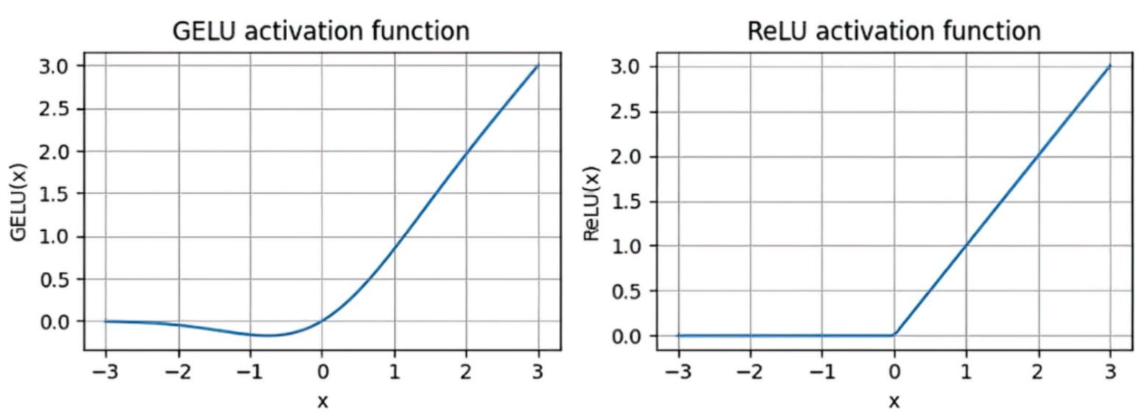

Figure 3-7 Curve of GELU and ReLU (source: [31])

The Dropout layer is used to avoid the model from learning the training data too well. And perform better on new data. The last linear layer brings the features back to the original input dimension.

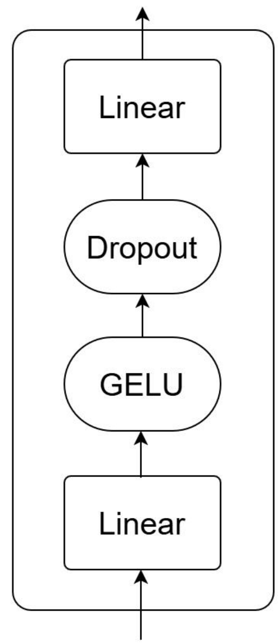

Figure 3-8 Structure of MLP decoder

Chapter 4

Market Data Analysis and Collusion Detection

This chapter outlines the full process of detecting collusive behavior. It starts with the collection and preparation of raw trading data, followed by an analysis of the key characteristics of electricity market data, an explanation of how input variables are constructed for the model, and finally, the introduction of an unsupervised anomaly detection model based on Transformer-AE.

4.1 Data Engineering

4.1.1 Data Collection

In this study, we downloaded the raw data from the official website of GME (Gestore dei Mercati Energetici). GME is the operator of the Italian electricity market (https://www.mercatoelettrico.org). It includes all daily bid submission records from each market participants. The data is provided in daily ZIP files, each containing structured XML documents, which makes them easy to parse in later stages.

To streamline the data collection process, we developed an automated script based on the Selenium framework[1], using Firefox's driver controller to simulate manual operations. The script automatically handles form clicks, date selection, and agreement confirmations, and incorporates appropriate wait times during the download phase to reduce the risk of interruptions caused by network fluctuations.

Given the potential market disturbances during the COVID-19 pandemic in 2019–2020, we intentionally excluded this period. Instead, we focused on the more stable window from 2021 to 2023, collecting complete market transaction data on a daily basis.

[1] By using selenium.webdriver.Firefox to control the browser and applying the find_element_by_name method to locate buttons and input fields on the webpage, we automated the process of transferring data "from the web to the local machine."

4.1.2 Data Composition

We downloaded daily data packages from the GME official website. They are all in ZIP format, within which multiple ZIP files are included. After decompression, we focus on the *"MGPOfferte"* XML file. This is bid submission data for each time interval in the Day-Ahead Market (MGP). In order to processing and analysis, all XML files are converted into CSV format and merged in time order.

Each record is bidding information of a specific time slot. Overall, the data-set provides a complete view of market participants' bidding behavior over the full 24-hour period of each trading day.

The Table 4-1 summarizes all the fields in the original files along with their corresponding meanings.

Fields	Description
UNIT_REFERENCE_NO	Unit code
OPERATORE	Registered name of the participant
BID_OFFER_DATE_DT	Date in YYYY/MM/DD format
INTERVAL_NO	Relevant period
QUANTITY_NO	Volume submitted by the participant
ADJ_QUANTITY_NO	Submitted volume, possibly adjusted by the system
AWARDED_QUANTITY_NO	Volume awarded by market
ADJ_ENERGY_PRICE_NO	Price possibly adjusted by the system
ENERGY_PRICE_NO	Price submitted by the participant
AWARDED_PRICE_NO	Price awarded by market
MERIT_ORDER_NO	Merit order calculated by market solution algorithm
SUBMITTED_DT	Time of submission
GRID_SUPPLY_POINT_NO	Grid supply point with which unit is associated
ZONE_CD	Zone to which the unit belongs
PURPOSE_CD	Purpose of bid/offer
TYPE_CD	It indicates whether the bid/offer is predefined or current
STATUS_CD	Status of bid/offer after market execution
PARTIAL_QTY_ACCEPTED_IN	Indicator of partially accepted bid/offer
BILATERAL_IN	It indicates whether the bid/offer comes from PCE platform

Table 4-1 Fields of data

4.2 Market Power Analysis Based on Italian Electricity Market Data

In order to examine possible instances of market power abuse or collusion within the electricity market, we analyzed transaction records from Italy's Day-Ahead Market. Data management and querying were handled using MySQL, while Python was employed to calculate various indicators. Our approach included the creation of long-term metrics to assess overall market concentration, alongside short-term measures aimed at finding sudden market shifts.

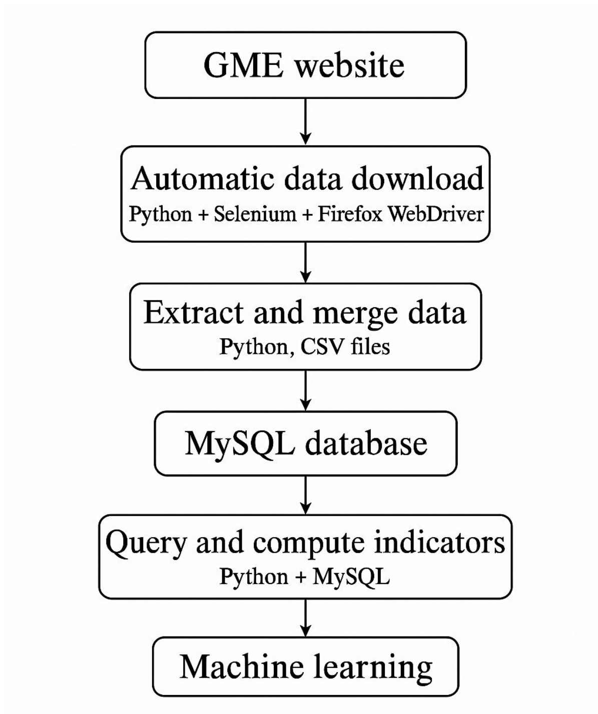

Figure 4-1 Flow chart of market power analysis

These indicators are derived from publicly available transaction data released by GME, covering multiple time points and dimensions of market participants. As shown in Figure 4-1. By analyzing the trends and cross-relations among these variables, we build a structure for finding potential risks in later stages of collusion detection.

4.2.1 Market Share

The table shows the combined market shares of the top 20 companies in the market over the past three years. Overall, a few leading companies control most of the trading volume: the top three companies alone have a combined share of over 42.1%, while the top 20 companies together account for 86.15%. Such high concentration indicates that companies with lower rankings have very limited influence over market prices. If control over market transactions is concentrated in just a few companies, it could disrupt market balance.

OPERATORE	Market Share (%)
G**A	15.07
E**A	12.28
A**A	7.44
EP**A	7.30
AX**A	5.33
EN**A.	5.05
ENI**A	4.61
SOR**A.	3.50
D**A	3.44
ENG**A	3.18
I**A	3.11
AL**A.	2.64
DO**A	2.44
T**A.	2.17
EG**l.	1.77
A**L	1.71
ER**A.	1.45
C**U.	1.33
DU**A.	1.30
IS**L.	1.04
Total	86.15

Table 4-2 Market share of top20 companies

The company has a big market share does not mean this company must have collusion, and small market share company also does not mean it must have no collusion. But it does pay out attention into them. In other word, it is risk. Because of complexity of electricity market, we need to eliminate all risks.

4.2.2 Long-term Indicators

Concentration ratio (CR_4 and CR_8)

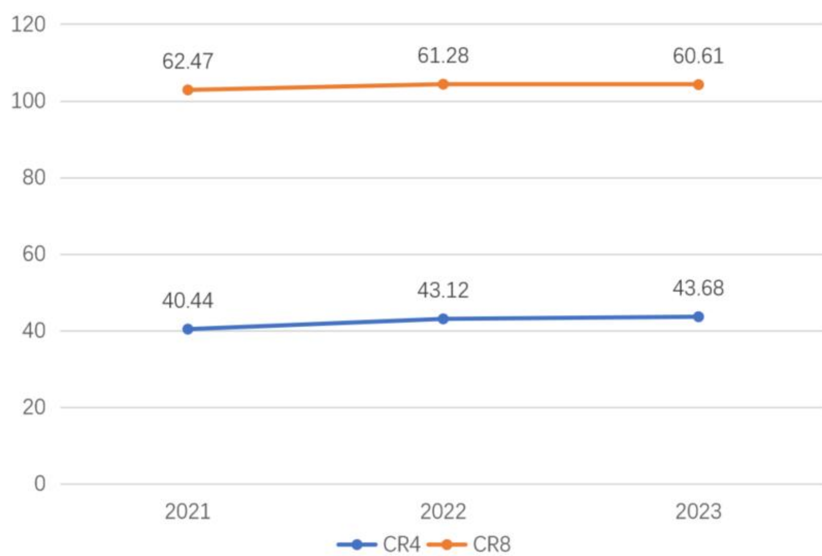

Figure 4-2 CR4 and CR8 in 3 years

Figure 4-2 above shows how the CR4 and CR8 indicators have changed over the past three years. It can be seen that the top four companies consistently hold more than half of the total market share. Although the CR8 indicator has slightly decreased, it still remains at a relatively high level. These data indicate that the overall market concentration hasn't changed significantly and continues to present a moderate concentration risk.

Herfindahl-Hirschman Index (HHI)

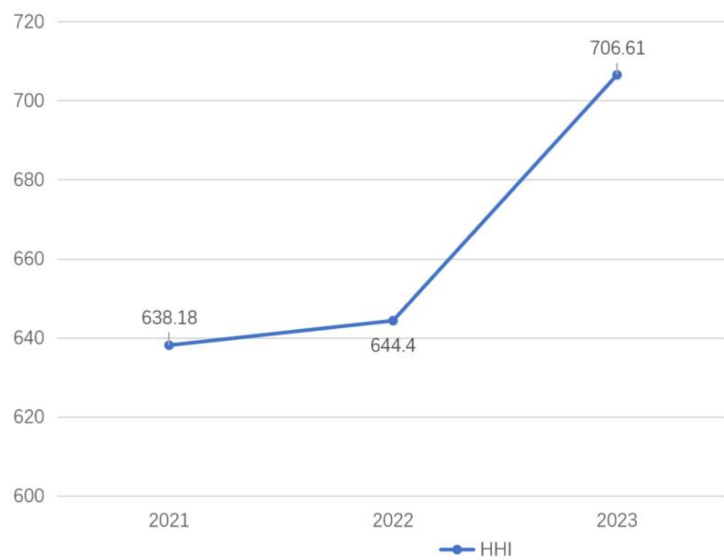

Figure 4-3 HHI scores in 3 years

The HHI index increased from 638.18 in 2021 to 706.61 in 2023, showing an overall upward trend in Figure 4-3. This indicates that market concentration has slightly increased. However, it remains at a low concentration level, suggesting that the market is still dominated by a few leading companies. Therefore, the overall market structure has not substantially changed.

Entropy coefficient (EC)

Year	EC	Ln(n)
2021	3.17	4.53
2022	3.20	4.55
2023	3.19	4.57

Figure 4-4 EC in 3 years

Both the entropy coefficient and market participation have increased slightly, but the improvements are quite small. As we learned from the previous sections, a lower entropy coefficient indicates higher risks. The entropy coefficient values over the three years are shown in Figure 4-4, means that the market may still have potential collusion risks.

4.2.3 Short-term Indicators

In this section, we calculated each short-term indicator in each time interval.

Risk	Date	Interval	CR4	CR8	HHI	EC	ln(N)
High	20221221	4	59. 53744767	70. 13054859	2493. 842251	2. 651681021	5. 204006687
	20221221	3	59. 20198304	69. 89453027	2463. 003809	2. 663919588	5. 192956851
	20221021	4	59. 72274628	67. 92830451	2458. 992475	2. 675958478	5. 187385806
	20221221	5	59. 05834313	69. 85075801	2442. 07739	2. 666867022	5. 204006687
Medium	20221008	4	55. 27291261	65. 67704529	1999. 764903	2. 842104642	5. 164785974
	20221203	2	54. 90121531	67. 37380358	1999. 361086	2. 823189665	5. 214935758
	20220303	2	59. 66153973	72. 08402013	1999. 261235	2. 730639295	5. 099866428
	20220828	6	57. 28136453	67. 69771309	1997. 465204	2. 812044427	5. 214935758
Low	20220117	21	49. 98002331	66. 07598311	999. 8839921	3. 11283471	5. 153291594
	20220422	17	48. 85887141	65. 06494766	999. 6091383	3. 201632504	5. 164785974
	20220120	10	49. 17263513	64. 8403602	999. 5603659	3. 142521652	5. 159055299
	20220422	10	49. 28845095	65. 29094321	998. 167754	3. 20640252	5. 176149733

Figure 4-5 Short-term indicators

The Figure 4-5 shows indicator values for high, medium, and low-risk levels at each time point. We can clearly see that the largest difference appears in the HHI index. Its ranges are from around 600 to about 2500. Almost doubling. This large difference is unusual. But there could be many possible reasons behind it, such as weather, unit maintenance, price changing, or even collusive behavior. At the same time, the other indicators show much smaller changes.That means each indicator has different sensitivity levels. With the current information, we cannot yet conclude whether this indicates collusion or not.

4.2.4 Section Summary

From the trading data of the past three years, the Italian electricity market generally shows a structure of "medium concentration." A few large companies control substantial market shares, leading to irregularly distribution. The CR4 and CR8 indicators have stayed at medium levels, while the HHI index remains slightly lower.

While long-term indicators are quite stable, the short-term data show large rolling during certain periods. This suggests that even though the market structure is generally stable, there might still be opportunities for short-term collusion. That also means that only statistical indicators cannot determine collusion. By using machine learning methods, we can better detect these unusual trading patterns.

Considering that collusive behaviors are usually hidden and the data is complex, our following analysis will mainly focus on power generation units with larger trading volumes and clear trading patterns. This will help the model more easily find relevant patterns and achieve better results.

4.3 Detecting Collusion with Machine Learning

In this section, we designed a framework. It combines statistical filtering and machine learning to detect potential collusion in the electricity market. Our method first uses statistical indicators to select specific units. After filtering, these units all have possibility in collusion. Next, we use machine learning models to learn the behaviors of these selected units. Maybe such as their bidding strategies and power generation at different times. Finally, we apply anomaly detection to find behaviors that different from normal patterns.We can say that these can be seen as potential signals of collusion.

Overall, we use "first filtering, then modeling, and finally detecting" to improve the efficiency of our analysis. It also effectively identifies unusual market activities without relying on labeled data.

4.3.1 Company Selection

Collusive behavior in electricity markets can appear in different forms. One type is horizontal collusion, occurring between different companies, and another is vertical collusion, happening among multiple units within the same company. In theory, collusion between companies is more damaging, but in fact, this type is harder to detect using only trading data because it involves complicated interactions like geographic location, grid management, and demand. Therefore, in this study, we mainly focus on detecting vertical collusion within a single company.

Using three years of market trading data, we selected the most relevant companies for behavior analysis. The selection was based on three conditions: total trading volume in the market, stability of pricing behavior, and overall market participation.

A. Total trading volume measures a company's influence in the market and represents how large its electricity supply is. The formula is below.

$$\text{TotalAwarded}_c = \sum_{t \in \mathcal{T}} q_{c,t} \tag{4.1}$$

B. The standard deviation of price offers shows the stability of each company's bidding behavior. A lower deviation suggests a more consistent internal strategy. The formula is below.

$$\text{PriceStd}_c = \sqrt{\frac{1}{|\mathcal{T}_c|} \sum_{t \in \mathcal{T}_c} \left(p_{c,t} - \overline{p}_c \right)^2} \tag{4.2}$$

C. Market participation is how active a company is during the entire period. And it including frequency that it bids and completes actual trades. This can help exclude companies irregular participation. The formula is below.

$$\text{Participation}_c = \sum_{d \in \mathcal{D}} \mathbb{I}_{c,d} \tag{4.3}$$

Based on these conditions, we need set specific thresholds for each company.For trading volume, companies had to be above 80% to ensure they have significant market influence.And pricing stability had to be better than the median value, indicating consistent bidding strategies. Finally companies needed to participate in at least one-third of all trading days to exclude infrequent traders.

OPERATORE	Total Awarded Sum	Price Std	Participation
G**A	68198880	10497.08	6147
E**A.	22868766	10582.57	6147
SOR***A.	15817963	10604.28	6147
ER**A.	6553861.5	10510.25	6147
I**L.	4714799	10375.88	6087
A**A.	3718676	10102.35	3058
T**H	2820041.5	10564.16	6147

Table 4-3 Selected companies

Companies that met these conditions generally have clear trading behaviors, high market participation, and significant market presence. In Table 4 - 3, we can find that SOR**A has the highest price standard deviation and high level of other two conditions. So, we chose **SOR**A** as the target company for our following sections and analysis.

4.3.2 Unit Selection

After choosing the target company, we further narrowed down the selection to specific generation units within the company. This was necessary because these units are different in capacity, frequency of operation, data quality and location. Some units have very few trading records, and others even have many missing data points, making them unsuitable for modeling.

Therefore, we applied two filtering conditions below.

First, we looked at time coverage, calculated as:

$$\text{Coverage}_u = \frac{|\mathcal{T}_u|}{|\mathcal{T}_{\text{total}}|} \tag{4.4}$$

We only kept units with trading records covering at least 90% of the total observation period. Complete data sets provide more valuable analysis.

Second, we considered behavioral activity, using two indicators:

$$\bar{Q}_u = \frac{1}{|\mathcal{T}_u|}\sum_{t\in\mathcal{T}_u} q_{u,t}, \quad \mathrm{Var}(P_u) = \frac{1}{|\mathcal{T}_u|}\sum_{t\in\mathcal{T}_u} (p_{u,t} - \bar{p}_u)^2 \tag{4.5}$$

We assessed units based on their average trading volume and the stability of their price offers. We set the threshold at the median value. In other words, we preferred units with frequent activities and noticeable changes because these types of units often provide richer information and maybe better prove potential collusion patterns.

UNIT_REFERENCE_NO	time_coverage	avg_quantity	price_variance
UP_CN**_1	0.9371	10.4558	17194.1170
UP_CN**_2	0.9288	10.2517	17363.1406
UP_FN**_1	0.9385	11.8887	17575.2632
UP_ME**_1	0.9040	32.9183	15436.9724

Table 4-4 Selected units

In the end, these four units are our object of study in Table 4-4.

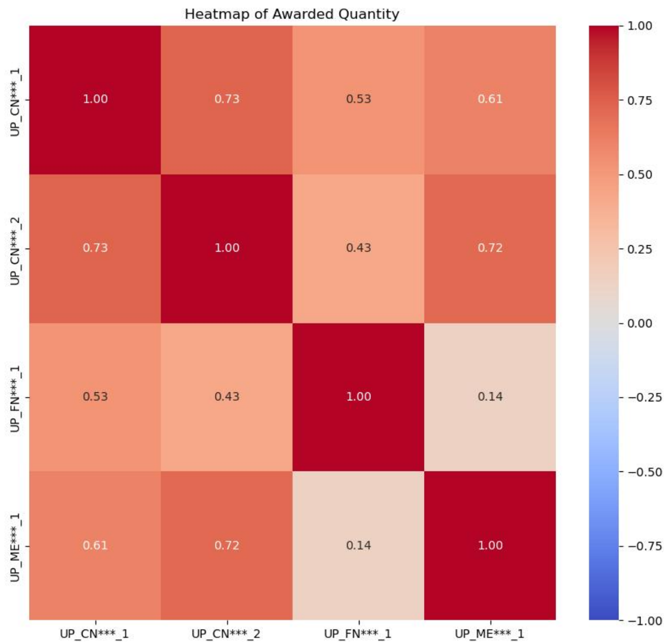

Figure 4-6 Heat-map of awarded quantity

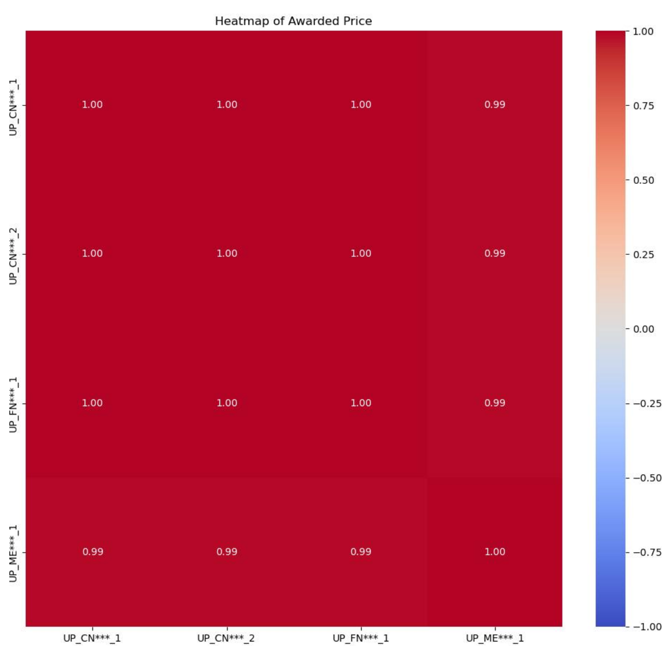

Figure 4-7 Heat-map of awarded price

The heat-map shown in Figure 4-6, Figure 4-7. They are the relationship between successful trading volume and bid prices for the selected units over the study period. From this figure, we can see that the trading volumes of different units show some similarities. As for bid prices, many units offer almost identical pricing. That might means a coordinated internal strategy within the company. This kind of behavior pattern deserves our attention.

4.3.3 Creating Input Features for the Model

To help the model better detect potential collusion, we built a set of input features combining economic and behavioral features. These features were collected for each specific time point (defined by both date and time period). We are going to create a feature matrix with timestamps as the index. The feature matrix is input for the Transformer-AE model.

We started by capture basic features, which is each unit's awarded price $p_{i,t}$ and volume $q_{i,t}$ at every time point. These can directly show us each unit's market behavior.

$$\text{Price}_{i,t} = p_{i,t}, \quad \text{Qty}_{i,t} = q_{i,t} \tag{4.6}$$

To capture dynamic behavior changes, we also calculated the rate of change for prices and volumes. We applied smoothing factors to reduce noise, making it easier for the model to spot significant trends. The formulas are below.

$$\Delta p_{i,t} = \frac{p_{i,t} - p_{i,t-1}}{p_{i,t-1} +}, \quad \Delta q_{i,t} = \frac{q_{i,t} - q_{i,t-1}}{q_{i,t-1} +} \tag{4.7}$$

Additionally, we introduced the price elasticity measure, which shows how sensitive traded volumes are to price changes. This can help us find if certain trades involve strategic adjustments[2].

$$E_{i,t} = \frac{\Delta q_{i,t}}{\Delta p_{i,t} + \epsilon} \tag{4.8}$$

We also built features based on collaborative behavior using the dynamic Eclat algorithm. This method identifies groups of units that frequently act together within a sliding time window. When a specific group of units appears simultaneously at a certain time point, we record this as a boolean collaborative behavior variable. This acts as a possible signal of collusion in the model input.

$$\mathrm{Supp}(X) = \frac{\mathrm{Count}(X \in \mathrm{Window}_t)}{w} \geq \mathrm{min\_support} \tag{4.9}$$

Finally, all these features were combined horizontally along the timeline, forming a complete feature matrix for model training and detection. An example of the complete input feature structure is shown below.

$$\mathbf{X}_t = [p_{i,t}, q_{i,t}, \Delta p_{i,t}, \Delta q_{i,t}, E_{i,t}, F_{X_1}(t), F_{X_2}(t), \ldots, F_{X_m}(t)] \tag{4.10}$$

Where m is the number of all units combinations.

[2] ϵ is a smooth factor, used to avoid division by 0. In our case, the value of ϵ is 10^{-4}

4.3.4 Data Processing

After creating the features, we needed to do some additional data processing. That is for making them suitable as input into the deep learning model. The general steps included cleaning unusual values, standardizing features, organizing data into a time-series format, and preparing batches for training. Only after completing these steps, we can feed the data into the model.

First, we deal with missing and extreme values. For example, some features had missing values at few time points. We filled these gaps using forward or backward filling. For extremely unusual values that were far from normal ranges, we set limits to keep them from negatively impacting the training. After cleaning, we standardized all the features using the StandardScaler to make it easier for the model to handle.

$$\hat{x}_t = \frac{x_t - \mu}{\sigma} \tag{4.11}$$

After StandardScaler , the mean of all features is 0, and the standard deviation is 1.

Next, we organized these features into sequences using sliding windows. Instead of isolated data points, the model could now see continuous behavior patterns over time. This sliding-window approach[32] can help the model better understand trends and changes.

We load the data in batches. This can save memory and make the training process more stable. Additionally, we disrupt the order of samples to avoid the model from memorizing patterns based on their order. Finally, we also use multi-core CPU parallel processing to speed up data loading and improve overall efficiency.

4.3.5 Model Training

Using the feature matrix we created earlier, we designed and built a Transformer-AE model to learn patterns in unit behavior sequences without needing labeled data. The overall structure of the model is in Figure 4-8.

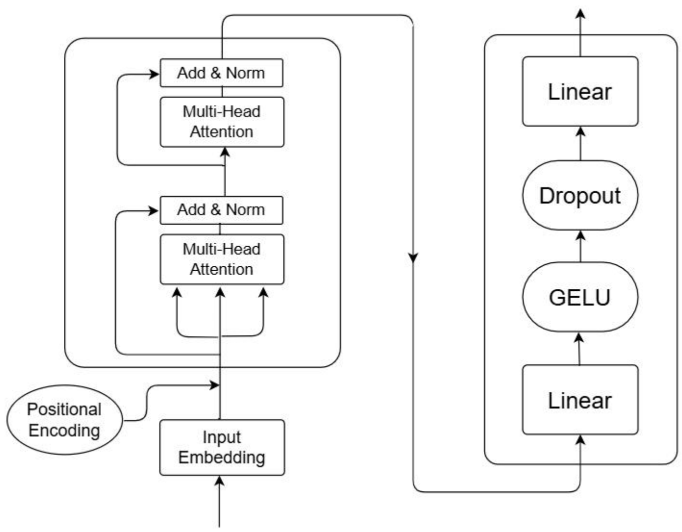

Figure 4-8 Structure of Transformer-AE

During training, we used an early stopping method to prevent the model from over-fitting. If the validation loss didn't improve for 5 epochs in a row, the training would stop automatically.

Figure 4-9 shows how the loss changed during training. We can see that the model quickly improved during early epochs and stabilized within the acceptable range, stopping at the right time.

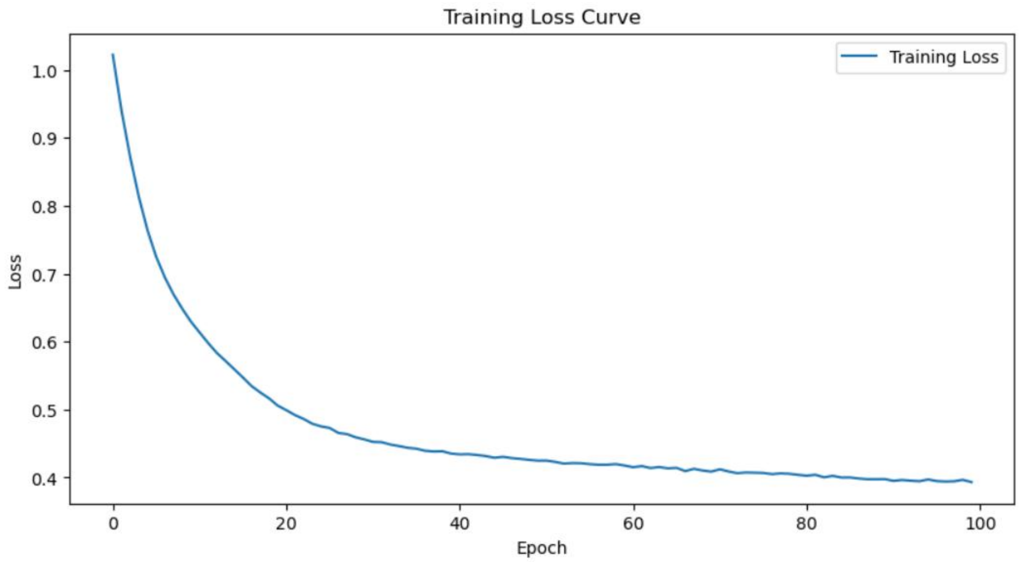

Figure 4-9 Training loss

4.3.6 Detecting Anomalies

After training the model, we analyzed the reconstruction errors for each time sequence to find unusual behavior. Because we used an unsupervised method, the training data had no labels. Therefore, the whole anomaly detection process relied on the model's ability to learn "normal" behavior patterns. In other words, we assumed that most unit behaviors are typical. The model learned these normal patterns. So whenever it encountered data significantly different from these patterns, it flagged them as potential anomalies. That can be seen as related to collusion.

First, we calculated the reconstruction error for each time sequence. That measures the difference between the model's predicted behavior and the actual data. Larger errors means the model learn it well, so that is a point of doubt. The formula is below.

$$RE_t = \frac{1}{D}\sum_{i=1}^{D}(x_{t,i} - \hat{x}_{t,i})^2 \tag{4.12}$$

Next, we use a dynamic threshold to select abnormal errors.

$$\theta = \mu_{\mathrm{RE}} + 3 \cdot \sigma_{\mathrm{RE}} \tag{4.13}$$

A dynamic threshold is better than fixed threshold, it calculates by mean and standard deviation. It can adjust based on the distribution of errors automatically. This can reduce false alarms. Especially when dealing with dynamic markets. According to the three-sigma principle[33], here we set the value of threshold coefficient to 3.

As we mentioned before, it is a abnormal if the reconstruction error bigger than threshold, otherwise it is normal behavior.

$$\mathrm{Label}_t = \begin{cases} 1, & \text{if } \mathrm{RE}_t > \theta \\ 0, & \text{otherwise} \end{cases} \tag{4.14}$$

Figure 4-10 shows the reconstruction error curve from the training phase, along with the actual anomaly periods detected by the final model on real-world data.

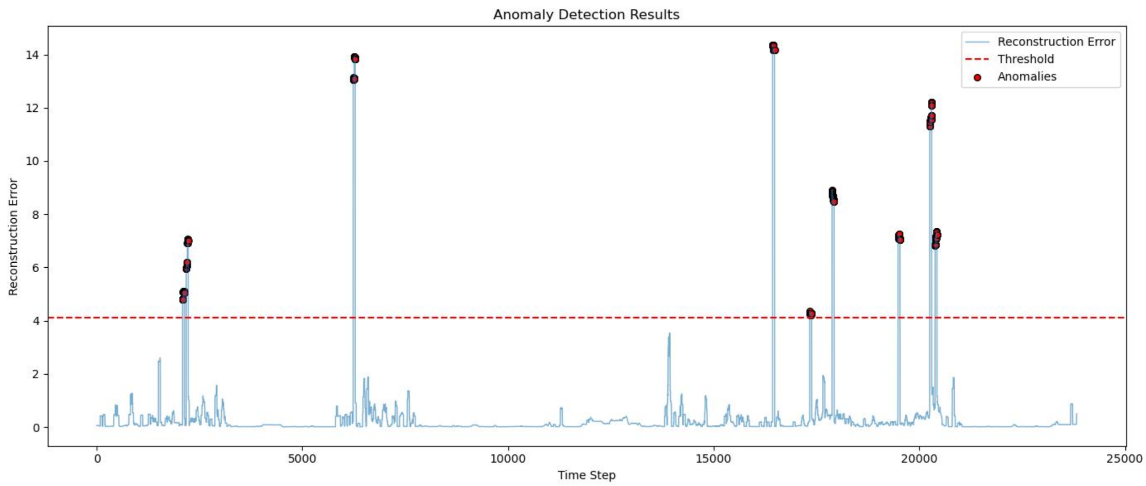

Figure 4-10 Anomaly detection results

Chapter 5

Case Studies

5.1 Case One

For better understand why these anomalies appeared, we visually examined the behavior of each generating unit during this period.

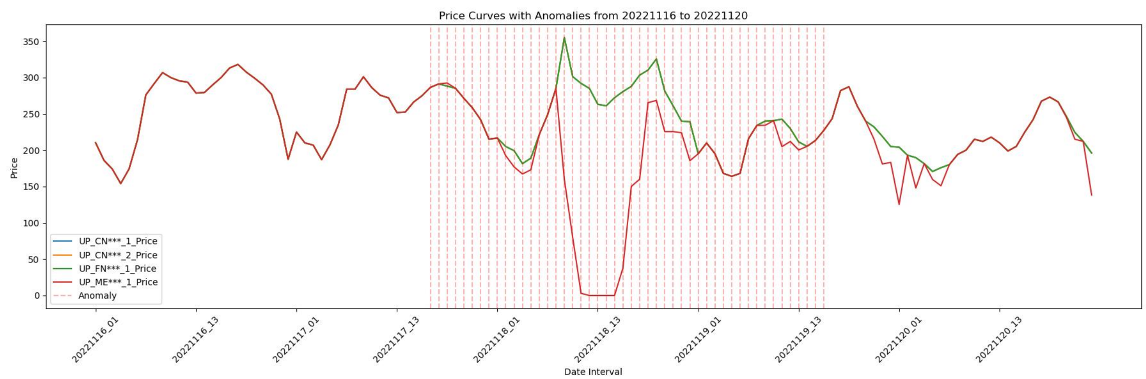

Figure 5-1 Price curves from 20221116 to 20221120

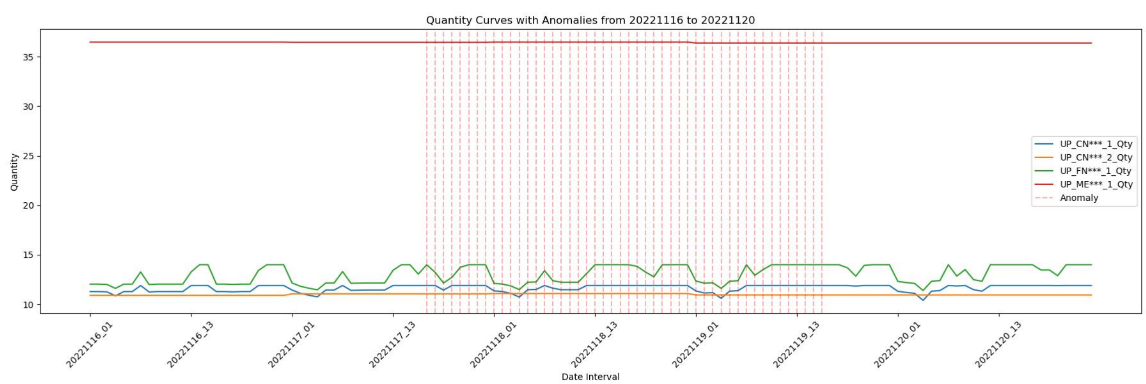

Figure 5-2 Quantity curves from 20221116 to 20221120

From Figure 5-1, it's easy to see that the unit UP_ME***_1 showed a sudden drop in its bid prices between November 18 and 19, maintaining a price near 0 €/MWh for several consecutive hours. At the same time, its generation output remained stable at around 37 MWh with almost no change. Compared to its historical records, this behavior clearly stood out, so the model labeled it as abnormal. Based on this result, we chose the period from November 16 to November 20, 2022, for a more detailed analysis of this unit's trading behavior.

Interestingly, this unit, which displayed clearly abnormal behavior, is from the Calabria region. During the same period, units from the NORD region, such as UP_CN***_1, UP_CN***_2, and UP_FN***_1, had some fluctuations, but their overall bid prices stayed within a normal range. This further highlighted the unusual behavior of UP_ME***_1.

Figure 5-3 Price curve in Calabria from 20221116 to 20221120 (source: [34])

However, to determine if this behavior was truly abnormal, we checked the market clearing price curve for the Calabria region in Figure 5-3[34]. We found that on November 18, there was about a six-hour window when market prices dropped to nearly 0 €/MWh before gradually returning to normal. This indicated that the entire region experienced a price decline, rather than this behavior being unique to a single unit.

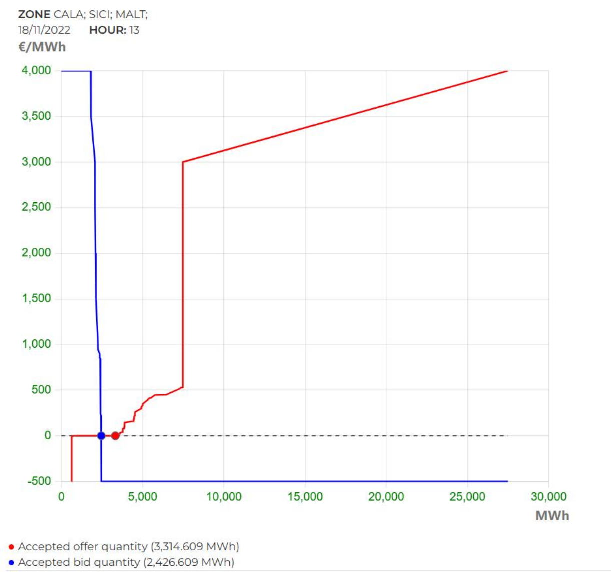

Figure 5-4 Supply and demand curves at 20221118 (source: [35])

Looking at the supply and demand curves, we also saw that during this period (Figure 5-4[35]) , Calabria had excess electricity supply, which typically pushes prices down to zero. For units like UP_ME***_1, bidding prices near zero was a reasonable according to these market conditions.

Overall, although this behavior was identified by the model as an anomaly. But it was actually a normal actions considering regional market conditions. This represents a typical false-positive case. And tell us it is important to have some additional market information.

5.2 Case Two

In the second case, the model detected an anomaly between January 17 and 19, 2023, as shown in the figure.

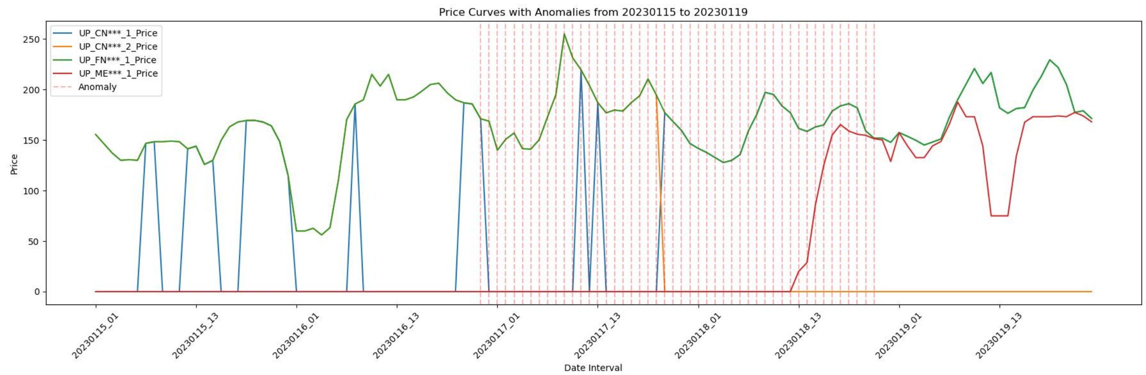

Figure 5-5 Price curves from 20230115 to 20230119

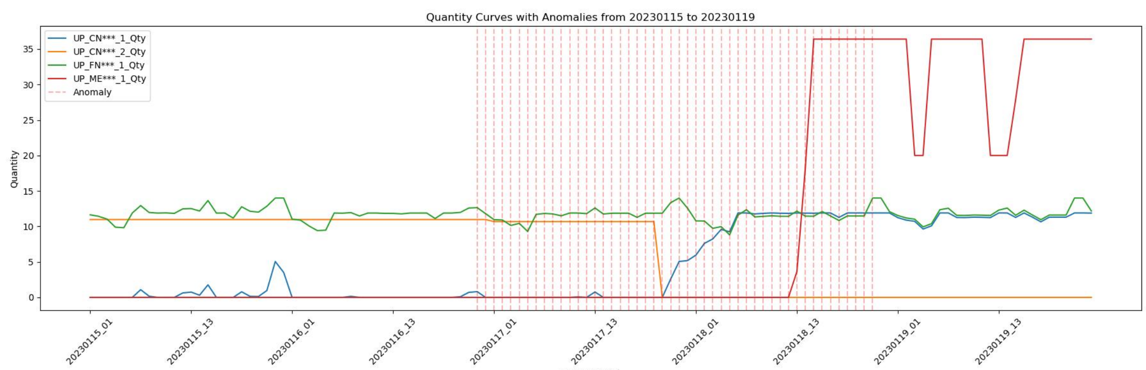

Figure 5-6 Quantity curves from 20230115 to 20230119

From Figure 5-6, we notice changes in behavior patterns for several units, but the unit UP_ME***_1 showed the most obvious anomaly, mainly around 1:00 PM on January 18.

First, let's look at UP_ME***_1. Before 1:00 PM, this unit continuously bid at 0 €/MWh and reported zero generation. So it is in a shut-down state. However, at 1:00 PM, it suddenly jumped to 37 MWh in generation. Its bid price also quickly increased. Because this sudden change differed greatly from its usual behavior, the model marked it as an anomaly.

Now, Let’s see what was happening across the entire market during this time.

Figure 5-7 Price curve in Calabria from 20230115 to 20230119 (source: [34])

Figure 5-7(source: [27]) shows that starting around 1:00 PM on January 18, the electricity price in the Calabria region quickly increased from about 20 €/MWh to nearly 100 €/MWh.

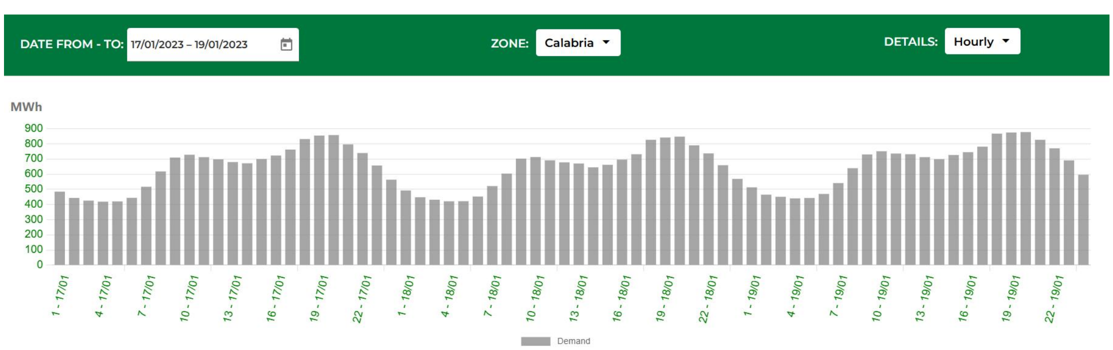

Figure 5-8 Demand in Calabria from 20230117 to 20230123 (source: [36])

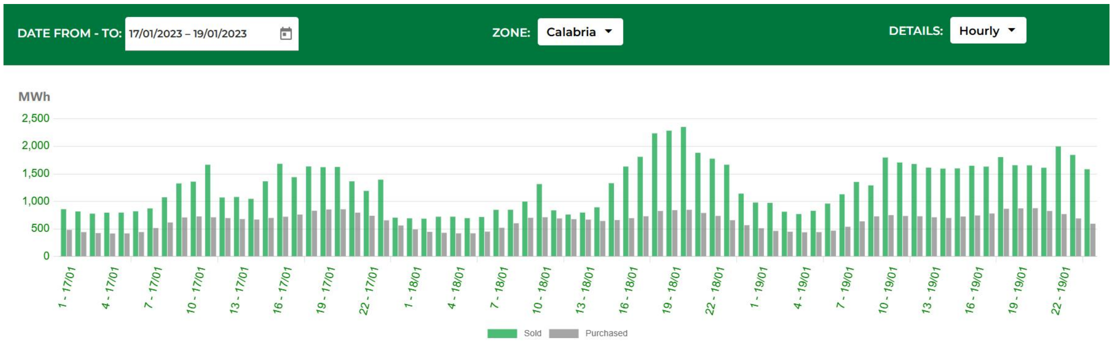

Figure 5-9 Sold/Purchased quantity in Calabria from 20230117 to 20230123 (source: [37])

Strangely, there was no noticeable change in the region's overall electricity demand[36] during this period, meaning that normal supply-demand logic couldn't explain the price jump.

Additionally, purchase data indicated that actual electricity purchases[37] during this period were clearly higher than usual, significantly above the same time in previous days. Combining this with the supply-demand curve (Figure 5-10[35]). We can find that the market was clearly oversupplied. That means available electricity exceeded what the market actually needed.

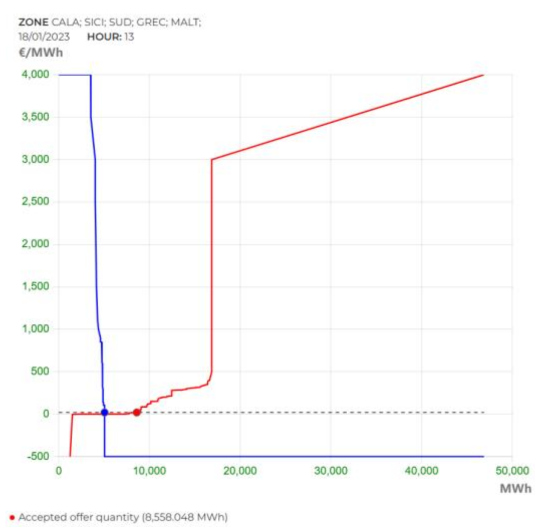

Figure 5-10 Supply-Demand curves at 20230118 (source: [35])

Under these oversupply conditions, the price would normally decrease. But instead, we saw an unusually sharp increase. This sudden rise might wasn't due to normal market rollings. It might have been caused by intentional actions, such as irregular grid dispatching or deliberate manipulation of marginal prices.

Therefore, while this case initially looked logical from as a single unit. The mismatch between prices and actual market supply-demand shows possible market manipulation or collusion. It represents an anomaly that deserves careful attention.

5.3 Case Three

In this case, the model identified anomalies occurring around March 25, 2023, as shown in the Figure 5-11, Figure 5-12

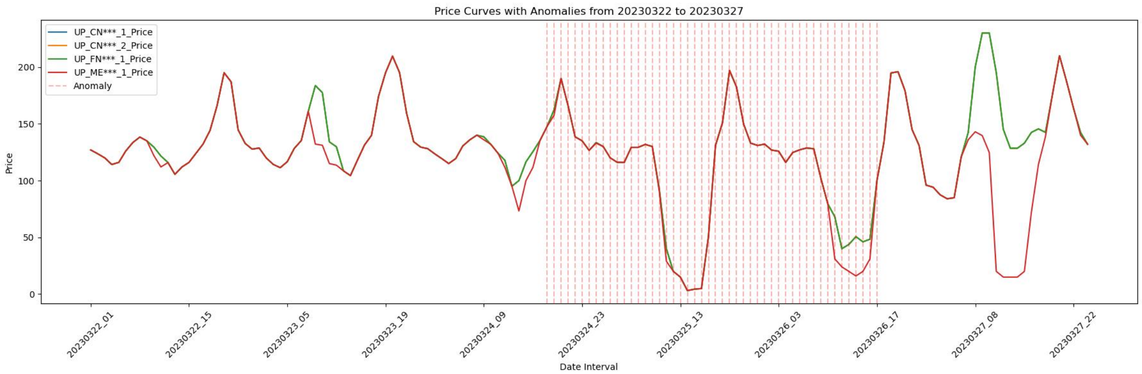

Figure 5-11Price curves from 20230322 to 20230327

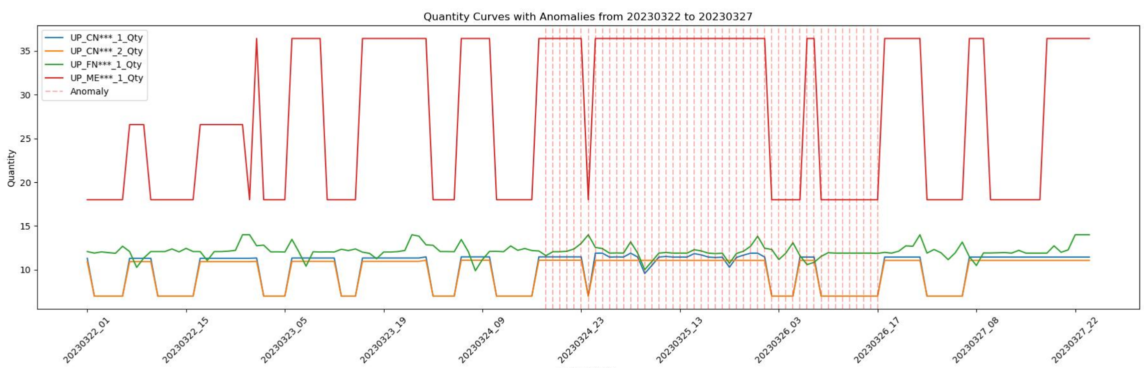

Figure 5-12 Quantity curves from 20230322 to 20230327

First, let's focus on a key unit in the Calabria region (the red line in Figure 5-11). Looking at the price data, this unit showed a noticeable change in its behavior around the anomaly period. Before the anomaly, both its price and output remained stable. However, during the anomaly, its generation suddenly stayed at full capacity continuously for several hours. If we look at the price alone, its bids matched the regional market trend. But this continuous full-capacity generation seemed unusual.

Figure 5-13 Price curve in Calabria from 20230324 to 20230326 (source: [34])

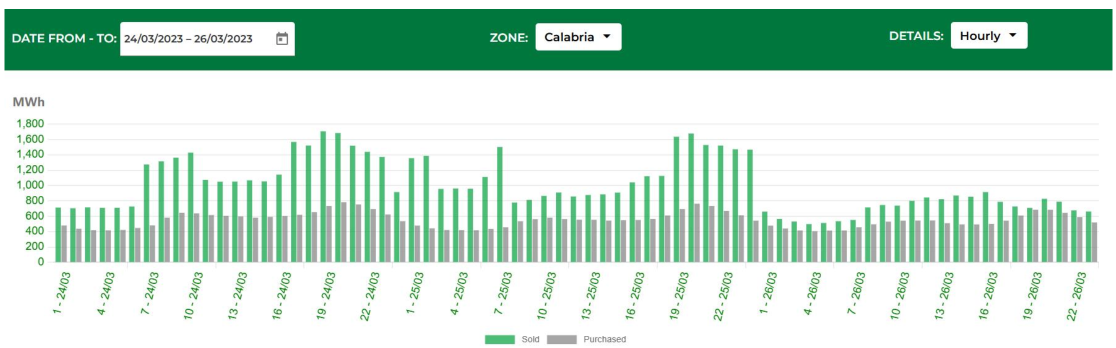

Figure 5-14 Sold/Purchased quantity in Calabria from 20230324 to 20230326 (source: [37])

Next, let's consider the three units marked in blue, yellow, and green in Figure 5-12, all located in the same region. Their bid prices were similar, but the unusual aspect appeared in their volume patterns. Normally, based on their naming, we'd expect the blue and yellow units to belong to the same station and show similar behavior. However, during the anomaly period, their outputs suddenly split, and instead, the blue unit closely matched the behavior of the green unit at several time intervals. The correlation changed in an unusual way. The internal teamwork broke down, and instead, it connected with outside parts. That seems pretty strange.

Figure 5-15 Price curve in North from 20230324 to 20230326 (source: [34])

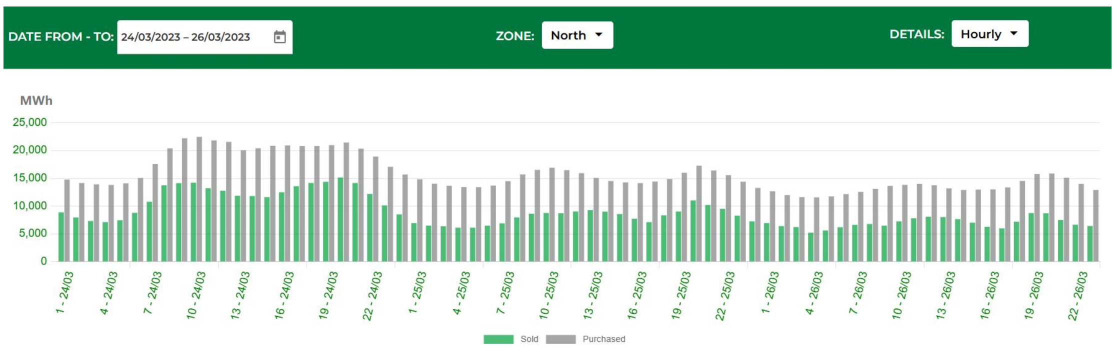

Figure 5-16 Sold/Purchased quantity in North from 20230324 to 20230326 (source: [37])

It's also important to know that these units are in different places. They are not directly connections by the grid's dispatch system. But, their price changes were still very similar. This makes it hard to explain their actions using normal grid control.

Because of these strange patterns, it makes sense to use anomaly detection. Besides, if we only look at regional market prices, the behavior might seem normal. But that might be a false positive case.

The key problem is many units act the same way, even though they are in different regions and stations. This makes people wonder if there are hidden deals or secret price plans. For market regulators, this kind of pattern should be monitored more closely.

Chapter 6

Conclusions and Future Directions

6.1 Conclusions

This thesis focused on finding and detecting possible collusion in electricity markets using modeling techniques. First, we introduced basic concepts of the electricity market, and then conducted a broad analysis based on public available trading data. By calculating several key market concentration indicators, we found that the market is medium concentrated. A few large electricity companies control a large part of market. So this is a chance for them to affect prices. Structurally speaking, this creates opportunities for collusive behavior.

From cases in chapter 2, the real collusion behavior is complex and happens in different ways. So we set up some conditions to filter useful units from major market players. This can help us make the model more focused and work more efficiently.

In our model, we used an unsupervised method with a Transformer Auto-encoder. We also added the Eclat algorithm to find patterns between different units. The goal was to detect unusual trading behaviors without using labeled data and to provide early warnings for possible problems. As shown in the case studies, our model successfully found several doubtful time periods. But in the end, final judgments still required manual checks external information such as market conditions and regional clearing prices.

Overall, this approach showed good sensitivity and flexibility in detecting complex behavior patterns, and providing early warnings about possible collusion. Although it can't directly confirm collusion, it gives a reliable support and a technical tool for investigations in the future.

6.2 Limitations

The first limitation is that the definition of "collusion" is not always clear. Different studies define collusion in different ways. Some approach use game-theory, and others focus on price coordination. In our study, we identified collusion by detecting abnormal trading behaviors. This method is more practical from a technical level, but it doesn't fully match the economic definition of collusion. In other words, our model provides an reminder rather than clear proof.

The second limitation comes from data availability. We used public available market data, which is quite comprehensive. However, important information like generation dispatch details, or records of communications between companies, is missing. This means our analysis only considers observable market behavior. The kind of information that can be seen as a collusion proof actually isn’t accessible with our current data.

Lastly, there’s the issue of determining whether an anomaly truly indicates collusion. Even if the model identifies behavior that clearly different from historical patterns, we still need context to make the judgments. Such as geographical location, supply-demand conditions at that time, and types of generating units. These cases still require human explanation. So, our model functions more like an early-warning tool rather than an automatic detection system.

6.3 Future Directions

In the future, there is still a lot we can do to improve this research.

First, if we can get real cases that were officially confirmed as "collusion", we can label these cases and use them to train our model. This would let us change from an unsupervised method into a supervised or semi-supervised approach. That would be improving accuracy. Having labeled data is definitely better for developing precise detection mechanisms, rather than relying on finding unusual behaviors.

Additionally, finding a clearer way to define "collusion" is extremely important. Right now, we mainly depend on anomalies for detection, but future studies could build systematic standard of behavior from multiple dimensions, such as intentional price manipulation or synchronized behavior patterns. This would help the model use not only algorithms but also clearer behavior rules.

Also, it's important to realize that collusion doesn't always happen inside one company. As we saw in our third case, it can happen between different companies or in different areas. So in the future, we can add regional factors, geographic and grid-dispatch information, or even weather conditions. These data can help us build a more complete model that better fits the real market situations.

[1] Nye, D. E. (1990). *Electrifying America: Social Meanings of a New Technology, 1880 – 1940*. Cambridge, MA: MIT Press.

[2] Joskow, P. L. (1974). Inflation and environmental concern: Structural change in the process of public utility regulation. *Journal of Law and Economics*, 17(2), 291 – 327.

[3] 国家能源局. (2024). 电力市场注册基本规则 [Basic Rules for Electricity Market Registration]. 国家能源局官方网站. Retrieved from https://zfxxgk.nea.gov.cn/2024-09/13/c_1212399683.htm

[4] Burlig, F., & Preonas, L. (2016). Out of the Darkness and Into the Light? Development Effects of Rural Electrification. Energy Institute at Haas, University of California, Berkeley. Working Paper No. 268.

[5] G. L. Lapini. "The power station in via Santa Radegonda." [Online]. Available: http://www.storiadimilano.it/citta/milanotecnica/elettricita/radegonda0.htm

[6] Enel Group. "60 years of access to electricity: Enel's journey since 1962." [Online]. Available: https://www.enel.com/company/stories/articles/2022/05/60-years-access-electricity

[7] Terna Group. "Bersani Decree: liberalisation of the electricity market." [Online]. Available: https://lightbox.terna.it/en/insight/bersani-decree

[8] International Energy Agency (IEA). "Feed-in premium for photovoltaic systems (Conto Energia V)." [Online]. Available: https://www.iea.org/policies/5210-feed-in-premium-for-photovoltaic-systems-conto-energia-v

[9] Imani, M. H. (2024). Empirical analysis of inter-zonal congestion in the Italian electricity market using multinomial logistic regression. Energies, 17(23), 5901. https://doi.org/10.3390/en17235901

[10] International Energy Agency (IEA). (2020). Power systems in transition: Challenges and opportunities ahead for electricity security. https://www.iea.org/reports/power-systems-in-transition

[11] Lerner, A. P. (1934). The Concept of Monopoly and the Measurement of Monopoly Power. The Review of Economic Studies, 1(3), 157–175. https://doi.org/10.2307/2967480

[12] Wolak, F. A. (2003). Measuring Unilateral Market Power in Wholesale Electricity Markets: The California Market, 1998–2000. American Economic Review Papers and Proceedings, 93(2), 425–430.

[13]Borenstein, S., Bushnell, J., & Stoft, S. (2000). The Competitive Effects of Transmission Capacity in a Deregulated Electricity Industry. The RAND Journal of Economics, 31(2), 294–325.https://www.jstor.org/stable/2601042

[14] Cramton, P., & Stoft, S. (2005). A Capacity Market that Makes Sense. The Electricity Journal, 18(7), 43–54.https://www.cramton.umd.edu/papers2005-2009/cramton-stoft-a-capacity-market-that-makes-sense.pdf

[15] ACER (Agency for the Cooperation of Energy Regulators). (2021). Market Monitoring Report 2020 – Electricity Wholesale Markets Volume. https://www.acer.europa.eu

[16] Hart, P. E. (1971). Ent.ropy and other measures of concentration. Journal of the Royal Statistical Society Series A: Statistics in Society, 134(1), 73-85

[17] Stigler, G. J. (1964). A theory of oligopoly. Journal of Political Economy, 72(1), 44–61.https://www.journals.uchicago.edu/doi/10.1086/258853

[18] European Union. (2012). Consolidated version of the Treaty on the Functioning of the European Union (Article 101). Official Journal of the European Union, C 326, 47–390. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12012E101

[19] Fabra, N. (2003). Tacit Collusion in Repeated Auctions: Uniform Versus Discriminatory. Journal of Industrial Economics, 51(3), 271–293. https://doi.org/10.1111/1467-6451.00199

[20] California State Senate. (2002). The California Electricity Crisis: Causes and Policy Options. California Senate Publications.

[21] European Commission. (2019). Antitrust: Commission opens investigation into possible collusion in electricity sector. Brussels: European Commission Press Release. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/ip_19_6216

[22] Mitchell, T. M., & Mitchell, T. M. (1997). Machine learning (Vol. 1, No. 9). New York: McGraw-hill.

[23] Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1), 59–66. https://doi.org/10.1080/00031305.1988.10475524

[24] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[25] Widenius, M., & Axmark, D. (2002). MySQL reference manual: documentation from the source. " O'Reilly Media, Inc.".

[26] Beaulieu, A. (2009). Learning SQL: master SQL fundamentals. O'Reilly Media.

[27] Hamidi, A., Hamraz, A. R., & Rahmani, K. (2022). Database Security Mechanisms in MySQL. Afghanistan Research Journal-Natural Science, 4(1).

[28] Zaki, M. J. (2000). Scalable algorithms for association mining. IEEE transactions on knowledge and data engineering, 12(3), 372-390.

[29] Hung, C. W., Li, W. T., Mao, W. L., & Lee, P. C. (2019). Design of a chamfering tool diagnosis system using autoencoder learning method. Energies, 12(19), 3708.

[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[31] Manning Publications. (n.d.). *ReLU*. LiveBook Wiki. Retrieved March 31, 2025, from https://livebook.manning.com/wiki/categories/llm/relu

[32] Vafaeipour, M., Rahbari, O., Rosen, M. A., Fazelpour, F., & Ansarirad, P. (2014). Application of sliding window technique for prediction of wind velocity time series. International Journal of Energy and Environmental Engineering, 5, 1-7.

[33] Walpole, R. E., & Myers, R. H. (2009). Probability and statistics for engineers and scientists. Prentice Hall.

[34] Gestore dei Mercati Energetici. (n.d.). *Zonal prices - MGP Results*. GME – Italian Electricity Market. Retrieved March 31, 2025, from https://www.mercatoelettrico.org/en-us/Home/Results/Electricity/MGP/Results/ZonalPrices#IntestazioneGrafico

[35] Gestore dei Mercati Energetici. (n.d.). *Demand Supply - MGP Results*. GME – Italian Electricity Market. Retrieved March 31, 2025, from https://www.mercatoelettrico.org/en-us/Home/Results/Electricity/MGP/Results/DemandSupply#IntestazioneGrafico

[36] Gestore dei Mercati Energetici. (n.d.). *Demands - MGP Results*. GME – Italian Electricity Market. Retrieved March 31, 2025, from https://www.mercatoelettrico.org/en-us/Home/Results/Electricity/MGP/Results/Demands#IntestazioneGrafico

[37] Gestore dei Mercati Energetici. (n.d.). *Volumes- MGP Results*. GME – Italian Electricity Market. Retrieved March 31, 2025, from https://www.mercatoelettrico.org/en-us/Home/Results/Electricity/MGP/Results/Volumes#IntestazioneGrafico