UNIVERSITY OF TURKU
Turku School of Economics

Detecting anomalies in wholesale electricity day-ahead market bidding data using LSTM-network.

Economics/Department of Economics

Master's thesis

Author:
Pekka Tuovinen

Supervisor:
Prof. Janne Tukiainen

26.4.2022

Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

Bachelor's thesis / **Master's thesis** / Licentiate thesis / Doctoral thesis

Subject: Economics
Author: Pekka Tuovinen
Title: Detecting anomalies wholesale electricity day-ahead market bidding data using LSTM-network
Supervisor: Prof. Janne Tukiainen
Number of pages: 66 pages
Date: 26.4.2022

In this thesis it is studied how neural networks can be used for anomaly detection in day-ahead wholesale electricity market trading data. Economics of electricity markets lays a foundation on detecting distinctive patterns in supply behavior reflecting market manipulation, such as economic and physical withholding of production capacity. The impact of market abusive supply behavior is studied on shapes of supply curves. A neural network model is used to provide score to measure stationarity of bidding behavior. An unsupervised machine learning framework for anomaly detection is set up by using a rolling window approach. 24 hours of high dimensional supply trading data is used as input to make prediction one hour ahead. Prediction errors of every individual hour are used as a time series of anomaly score, which is thoroughly analyzed in the light of signs of market manipulation based to the literature of electricity market economics. The study is conducted on two years of anonymous aggregated day-ahead trading data collected by The European Union Agency for the Cooperation of Energy Regulators received from Energy Authority of Finland.

Idea is to fit neural network to the data to estimate how supply curve of an hour would look like conditional on 24 previous hours and external variables. Neural networks are used for the estimation as they are capable of modelling non-linear spatial dependencies in the data. LSTM model is further chosen because it is designed to handle long term dependencies in the data. If the prediction errors are low enough on average, it can be assumed that the model can capture stationary behavior in the data and outliers can be assumed to result from changes in data generation process. If model can predict supply curves well enough on average, large prediction errors can indicate that something unexpected has happened in the markets. LSTM-model is trained to make rolling window predictions using 5-fold walk forward validation approach, where chronological order of the data is maintained to mimic real life prediction scenario. Early stopping is used to prevent overfitting. Hyperparameters are chosen via grid search likewise using 5-fold walk forward validation.

Two major distinctive types of supply behavior are identified from the literature, economic withholding and physical withholding. Their impact is studied on supply curves and is paid attention in the analysis of anomaly score. Mean absolute error of individual hour is chosen for anomaly score, which is referred as h-MAE. Performance of the model is compared to one used by Guo et al. (2021) in similar function of predicting supply curves. Method is promising in detecting out-of-ordinary supply curves, based on thorough statistical survey of the results and brief qualitative survey, in which a confirmed market violation was detected, as well as erroneous period in the data. Linking market manipulation to the anomaly score directly proves difficult. However, the method offers a noteworthy possibility to surveil and study supply curves in day-ahead market as it benefits from enormous amount of high-dimensional data and is capable to take account the spatial and temporal non-linear relations of the supply curves.

This Master’s thesis was done in affiliation with Energy Authority of Finland.

Key words: Unsupervised machine learning, anomaly detection, electricity markets, day-ahead market, long short-term memory, market manipulation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Electricity markets and surveillance

Historically electricity markets around the world started out as regulated monopolies of large vertically integrated utilities. All sectors of electricity markets were typically integrated under ownership and operation of one such actor. The early model of regulated regional monopolies was consequence of large investment costs in electricity production and distribution, and physical qualities related to electricity as commodity that led to natural monopolies. Regulated monopolies have since given way to reconstructed, or liberalized, market design around the world, where disintegrated market-design and market-oriented mechanisms are set up to induce competition (Brown & Olmstead 2017). Around the world markets have been reconstructed from regulated monopolies to disintegrated monopoly design. Finland's electricity market was opened to competition also gradually from 1995 to 1998 after the Electricity Market Act was passed. Finland joined in integrated Nordic power markets in 1998, which was the evolved into Nord Pool of today that operates multiple markets in 20 countries including Baltic and UK. Reconstruction of electricity markets in Finland meant that retail, distribution and production of electricity were separated and opened for competition. Electricity producers started to sell production in stock markets provided by Nord Pool.

Although many electricity markets today have been opened for competition, some inherent qualities in electricity production have not changed. Investments in most profitable generation technologies still remain large and are in scope of only the dominating producers. Network congestions and bottlenecks can also result in monopolistic traits in electricity production locally. Even thought Nordic markets were among the first to open up for competition and first to establish international power exchange, market concentration in electricity production was still high even in 2009. There is ongoing dispute over best methodology when it comes to assessment of market concentration and views on its impact differ. (Hellmer & Wårell 2009).

Electricity markets differ from most markets in the absence of possibility to store production. Even though the technology for storing electricity is constantly improving, efficient long term electricity storages is not yet viable option for large scale usage. This

leads to balance required between consumption and production of electricity at all times, which leads to high intertemporal supply variation. In addition to the high intertemporal variation of electricity supply, the demand of electricity is very inelastic in the short run. One reason for the inelasticity of demand is retailers protecting their customers from price variation, which leads to indifference in consumption decision among the final consumers (Borenstein et al. 2002). Properties of electricity as a commodity and concentration of electricity markets, make the market vulnerable to market manipulation. As such, electricity markets are strictly regulated. Market manipulation is prohibited in the regulation of the European Parliament and of the council of 25 October 2011 on wholesale energy market integrity and transparency (REMIT).

Reliability of transparent and functioning electricity market rely on surveillance. Surveillance of electricity markets does not come without challenges. The number of transactions and actions taking place daily is enormous, as there are multiple parallel markets operating simultaneously. Intertemporal variation and high dimensionality brought by the market mechanism makes statistical inference and manual surveillance difficult. Trading activity in electricity markets is increasing due to increasing utilization of algorithm trading (Epex 2022). Also, in addition to intertemporal variation brought by factors like weather, time of year and time of day, the underlying circumstances in the markets are constantly changing, new production and consumption enters and exit the market, consumption profiles change, and new technologies emerge. Surveillance of electricity markets need to adapt as going through the enormous daily data feed manually can quickly prove impossible. Automatic surveillance is called-for. Problem with automatic surveillance is the intertemporal variety, it is difficult to label transactions suspicious based on some threshold, when volume of demand and supply are highly seasonal and circumstances like weather and price of fuel impact them. Therefore, a more flexible way is needed. Deep learning algorithms are widely used for anomaly detection across many demanding fields (Lindemann et al. 2015). Such applications could also benefit surveillance of electricity markets, since high dimensional seasonal data with non-linear relations restrain inference by parametric models and traditional time series analysis.

1.2 Research question and structure of the thesis

The first research question of this thesis is to find out if neural networks can be used in unsupervised anomaly detection to detecting abnormal events in day-ahead markets. Second research question is, do the abnormal events indicate market manipulation? Strategy is to predict day-ahead supply curve based on supply curves that occurred in previous 24 hours and assess the hours that result in high prediction error. Study is conducted on aggregated anonymous day-ahead market supply curves of Finland's price area received from Energiavirasto. Labelling hours strictly anomalous or non-anomalous is not the focus of this thesis, but to examine if the method is capable to capture elements in the supply curves that could indicate market manipulation, erroneous orders or unexpected abnormal events and inspect the results. In chapter two, market design and regulation are described as well as economic theory related to electricity markets and market power to shed light on how market manipulation should show in supply curves. In economic approach, first optimal bidding in perfect competition is defined, and after that, motivation to differ from perfect competition is examined in light of monopoly profits. In chapter three, anomaly detection framework is introduced as well as literature survey on prediction of supply curves, in which the framework relies on. In chapter four method is explained and long short-term memory network is introduced. Also, the details of model selection are reported. In chapter five the results of model selection and model fitting are reported with examination of the data distribution. In chapter six the results are examined both qualitatively and qualitatively and discussion of the shortcomings and successes of the study and perspectives on future directions of developing the method are provided. The thesis ends in summary in chapter seven, conclusion.

2 Electricity market description

2.1 Day ahead electricity market

2.1.1 Price formation and spot price

In electricity markets, there are two important centralized market processes to be considered, day-ahead and intraday markets. In addition, there are also future markets and regulatory power markets, which are left outside the scope of this thesis. Intraday and day-ahead markets operate in two stages. First, before the actual dispatch, bids and asks are placed to the day-ahead market, which form step functions. Market operator calculates optimal dispatch over every hour of the subsequent day, and the market participants are paid according to the provided forecast price and forecast dispatch. All supply bids summed horizontally form market supply bid curve, the market demand curve is formed the same way by summing all the demand asks. The day-ahead price is determined by cross-section of market supply bid curve and demand ask curve. From now on, the curve that is the sum of supply bids is referred to as market supply curve, and respectively the sum of demand asks is referred to as market demand curve. Market participants provide bids and asks for every hour of the subsequent day, so there are 24 market supply and demand curves for each day in day-ahead markets resulting in day-ahead, also called spot, price for every hour of the day. In figure 1 cross section of market supply and demand is illustrated. Step-like shape of the curves results from bidding types available on day-ahead markets elaborated in the chapter 2.1.3 and the minimum operating level of certain electricity generation technologies. The most expensive technologies incorporate higher marginal costs and thus place higher in the merit order of supply bids. It is typical for market supply curve to present a hockey-stick like shape, less volume is offered at higher prices. Less volume offered at higher prices is also consequence of rare need to dispatch generation technology that has high marginal cost, so in order to cover the fixed costs the production of high marginal cost technology must be offered to the market with a price exceeding the true marginal costs. Most of the variance in electricity price takes place in the midrange of the merit order of supply (Ziel & Steinert 2016). It is important to note, that renewable energy that does not incorporate variable costs in their production, such as solar and wind production, move the supply curve to the right when available, since without variable costs it is always financially beneficial to offer their capacity to the coupling at any non-negative price. In practice this is achieved by offering supply at

minimum price allowed. In figure 1 the shapes of supply and demand curves are illustrated.

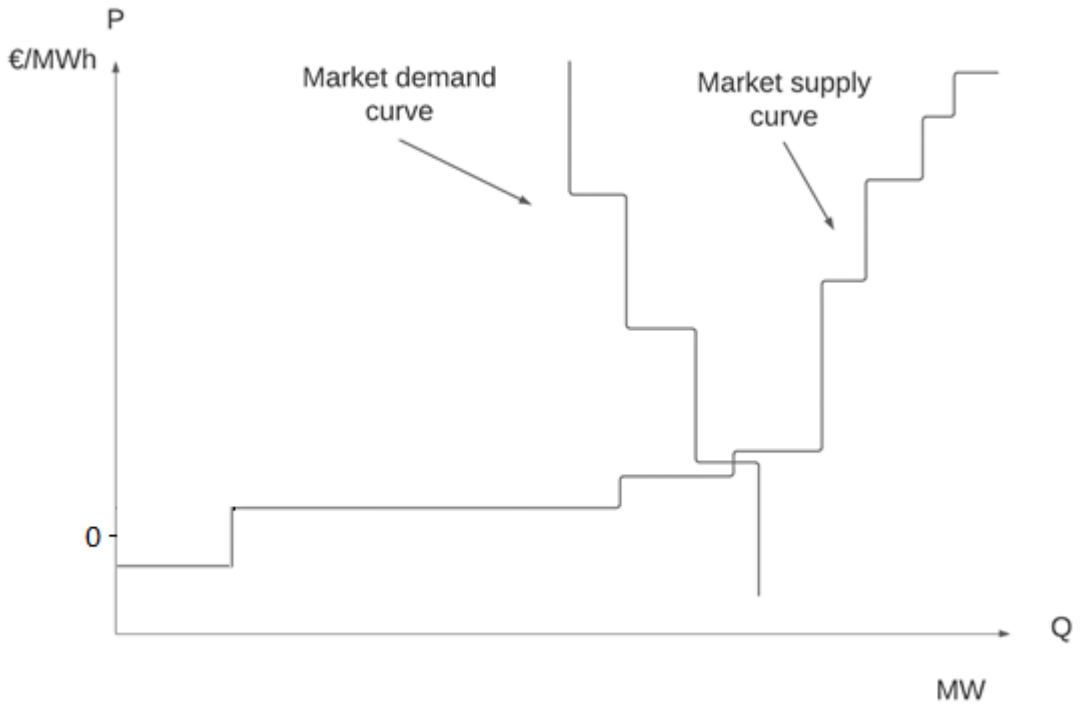

Figure 1, Cross-section of market supply and demand curves determine the spot price for each hour of the subsequent day.

Intraday market occurs during the day of actual dispatch, the market operator uses bids and asks placed to the intraday market to couple supply and demand conditions resulting in real-time intraday price and dispatch. Day-ahead electricity markets can be thought a as forward market. If the price and quantity for market participant in the day-ahead market are P^{DA} and Q^{DA}, the revenue from day-ahead market is $P^{DA}Q^{DA}$. In addition, market participant can sell the difference in real-time and day-ahead dispatch for $P^{ID}(Q^{ID} - Q^{DA})$ The total revenue that market participant receives for supplied energy can be thus expressed by

$$P^{ID}Q^{ID} + (P^{DA} - P^{ID})Q^{DA}$$

The equation above describes the relationship between intraday price and supply behaviour, fundamentally only the intraday price determines the revenue and day-ahead market can be considered to operate as a short-term hedge for intraday price. In day-ahead market, individual market participant delivers its bids to the wholesales market operator resulting in market participants supply curve. It is worth highlighting that supply curve of an individual market

participant does not represent marginal costs, but the merely the bidding behaviour. (Biggar et al. 2014.)

2.1.2 Nord Pool Day-ahead

The scope of this study is limited to the day-ahead market of Finland. Finland is part of integrated Nordic and Baltic energy market where the market operator is Nord Pool. In day-ahead market participants place sell and buy bids for every delivery hour of the next day. Bids are placed in a two-hour time window between 10:00 CET and 12:00 CET on the day before delivery. Bids are placed on a grid that consist of price steps and individual delivery hours. (Nord pool 2021a.)

Day-ahead trading is based on four types of orders. Single hourly orders, block orders, exclusive groups and flexi orders. Market participants may derive other types of orders from four basic order types.

Firstly, single hourly orders mean that participants specify the buy or sell volume for each hour. Single hourly orders must be defined with at least two price steps, which means they can be either price dependent or independent, since participant can set the price interval to match the minimum and maximum market prices. According to Nord Pool, the largest share of the day-ahead trading is based on single hourly orders.

Secondly, block orders are set with specified volume and price for several consecutive hours. Regular block orders must be fully accepted or rejected, which means that all the hours the contract covers shall. Block orders are accepted when the average day-ahead price is lower (higher) than sales (purchase) block orders over the whole timespan of the order. Block orders can be linked so that individual block order can be made dependent on acceptance of another block offer. Additionally, volume of the block order can vary over the time span.

Thirdly, exclusive group means a cluster of block orders out of which only one can be activated

Lastly, Flexi order are block orders with maximum timespan of 23 hours. Flexi sales orders are used by companies for example to sell power to the day-ahead market by closing industrial processes. (Nord Pool 2021a.)

The shape of the market participants supply curve for each individual delivery hour is formed by combination of three types of sell orders from those described above, exception is block orders, which are dependent on intertemporal conditions. Furthermore, the supply curve takes its form as a collection of offered volumes on corresponding price for each delivery hour. In Nord Pool, pricing of offered load is limited between upper and lower thresholds of 3000€ and -500€. Block offers do not have effect on the shape of the supply curve because of their price independent nature, the whole block is either accepted or declined. Therefore, accepted block offers add to the volume of each price-volume pair, simply shifting the market participants supply curve to the right. To illustrate the daily market coupling, eight aggregated anonymous market supply and demand curves from Finnish price area are presented from one day in figure 2.

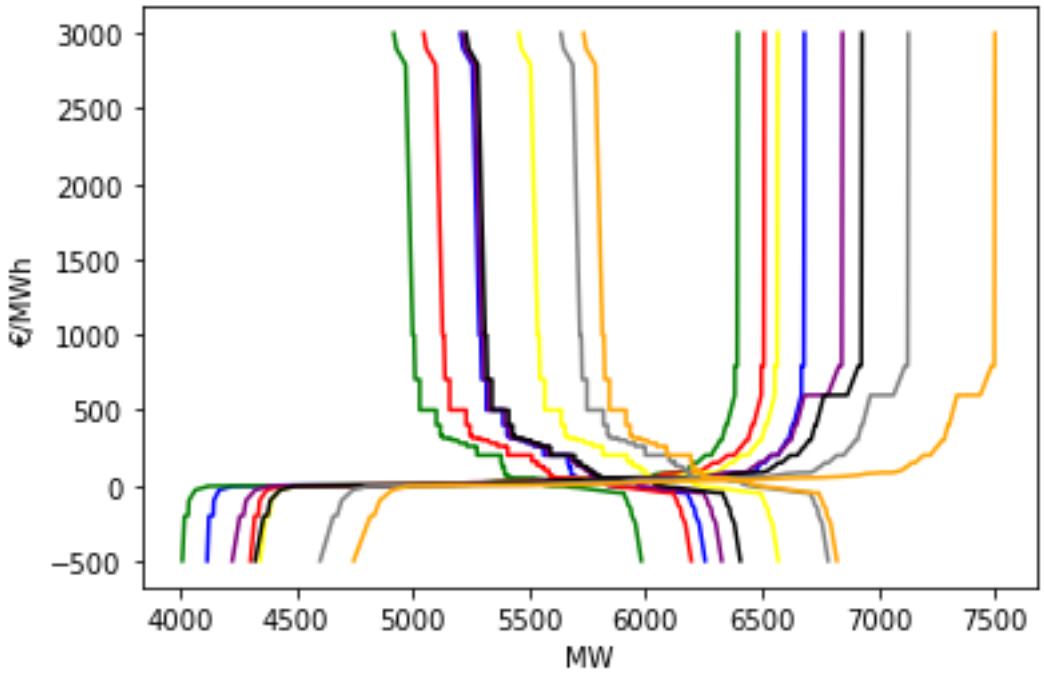

Figure 2, Aggregated market supply and demand curves (Nord Pool 2021)

In figure 2, same colour represents supply and demand curve of the same hour. The shape of both curves is determined by fundamental conditions in the market. Demand of energy is inelastic during short period like a day, but changes in volume which shifts the position of the demand curve along the horizontal axis. On the other hand, supply conditions have greater impact on the shape of the supply curve. For example, the amount of available electricity generation without variable costs, such as wind shifts the curve to the right on horizontal axis as there are more price-independent supply offers. Both supply and demand curves show a great deal of variance within one day. In this this thesis later on,

bids will be divided to 15 intervals: from -500 € to -10 €, from -10 to 0 €, from 10 € to 100 € with increment of 10 €, from 100 € to 200 € and from 200 € to 3000€. In figure 3 the supply curves from figure 2 are represented after division to price intervals, or price bins.

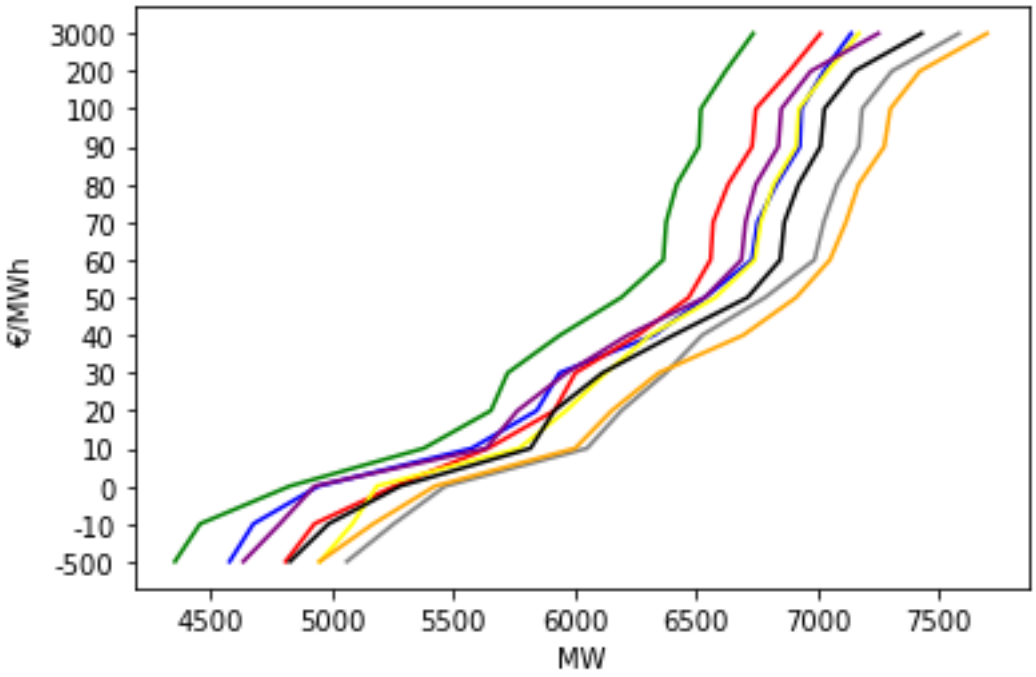

Figure 3, Aggregated supply curves with bidded volume divided to price intervals.

2.2 Perfect competition and market power

2.2.1 Perfect competition

In perfect competition conditions, each market participant has incentive to bid according to their marginal cost of supply or marginal value of consumption. In perfect competition market operator treats every supply and demand curve as truthful representations of underlying marginal costs and there are adequately many firms taking part in the market with equally low market share, so no firm on supply or demand side has direct influence on the market price. Furthermore, there are adequate amount of generation technologies utilized by many firms. The system marginal cost (SMC) is determined by sum of supply and offer curves, since no participant can gain profit by differing their supply behaviour from their marginal costs. SMC in this chapter refers to theoretical aggregated marginal cost of the electricity supply and is not to be mistaken with short-run marginal cost.

System price in perfect competition shall be equal to SMC and is determined by demand. In figure 4, the bidding behaviour in perfect market conditions is illustrated as well as speculative non-rational supply curve.

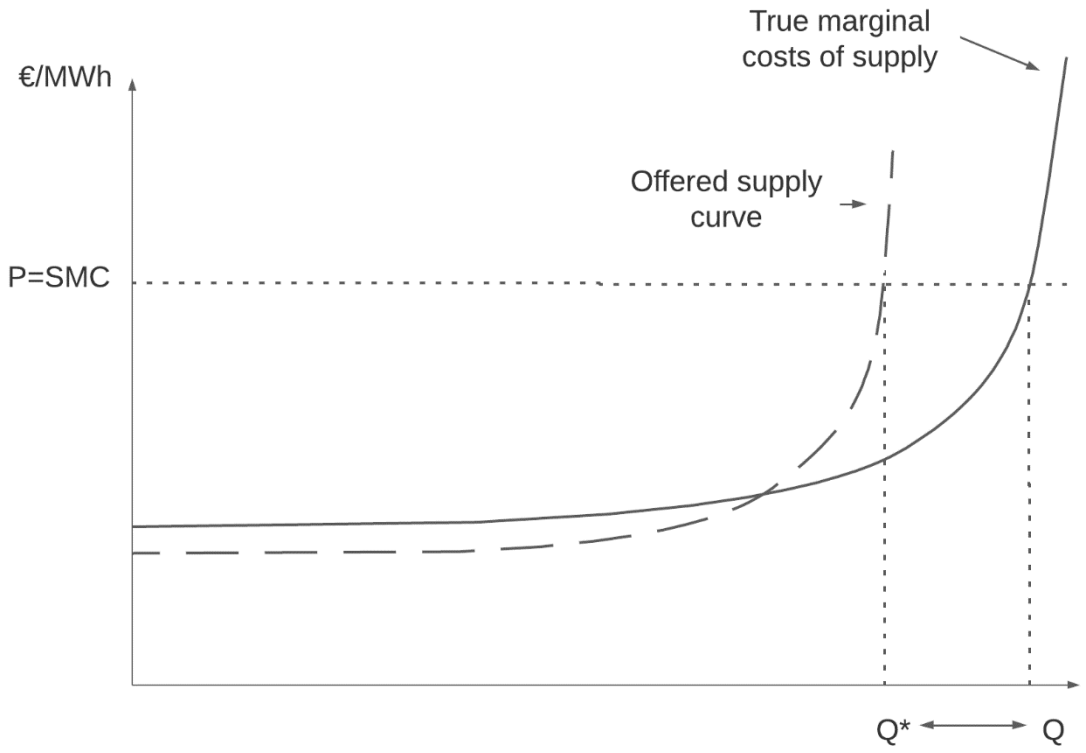

Figure 4, In perfect competition it is not beneficial for supplier to bid outside their marginal costs

If market participant expects a price P and bids their available volume at a price above their marginal costs, their quantity of supplied energy accepted in the coupling at that price is less than by offering at marginal cost, thus shifting from Q to Q*. Therefore, there is a strong incentive to offer all volume available at that marginal cost. On the other hand, there is no incentive on offering volume at a price below the marginal cost since it returns negative profits. (Varian 2020 & Biggar et al. 2014.) The perfect market conditions for electricity market are very strict and somewhat unrealistic. In truth, the markets typically are concentrated as small number of power companies control most of the generation capacity (Tangerås & Mauritzen 2018). Instead of perfect competition, circumstances in electricity market rather seem to show traits of oligopolistic competition. Reasons behind concentration in electricity markets can be economics of scale, political and economic barriers making entering the markets difficult, and transmission bottlenecks limiting trade. Demand of electricity is considerably inelastic, due to popularity of fixed price contracts and the lack of possibility to shift consumption during high prices. The stock of

generation assets is fixed in the short-run supply, electricity producers typically can only increase their output during the time of high demand to a limited scope, resulting in small and expensive suppliers might having a high influence on the market price. Structural, unexpected and expected network constraints can reduce the size of the area over which the suppliers can compete raising the scope of localised market power. The suppliers are continuously in interaction with each other in the market process. The repeated nature of the coupling process gives opportunities for suppliers to develop reputations, give signal about their intentions and to evolve towards co-operative or collusive arrangements. (Hesamzadeh 2014.)

2.2.2 Market power in electricity market

Across economics literature there are vast number of definitions for market power. One simplest definition for market power used broadly in economics is to define a firm that is not a price taker to have market power. It is broadly accepted in economic literature that a firm without possibility to influence market price by its actions does not have market power. Following this definition, in electricity market, an energy supplier can be stated to possess market power if it can influence the market price by varying its rate of supply. Thus, a firm altering its bids in a way that is deliberately designed to alter the wholesale market price can be defined as exercise of market power. (Hesamzadeh 2014.)

In chapter 2.1.2 incentives of the market participants were demonstrated to lead to bidding at their marginal costs in conditions of perfect competition. As discussed, the condition does not represent real day-ahead markets and therefore incentives to exercise market power cannot be ruled out. Oligopolistic market power in electricity markets is typically modelled by either Cournot model (Neuhoff et al. 2005) or supply function equilibrium model (Klemperer & Meyer 1989). Both models assume that firms make bidding decisions assuming that the supply function of competitors remains fixed. Both models are capable to explain almost the same fraction of price variation in Germanys electricity markets. However, Cournot model is preferable in short-term analysis because technical details are easier to include in the analysis, and supply function equilibrium is preferable on long-term modelling where sensitivity of calibration parameters can become hinderance (Willems et al. 2009). Regardless of the model used, the exercise of market power in bidding is done with the same measures. As defined in chapter 2.1.2, if a market participant exercises market power, by definition, it influences on price through supply

behaviour. That is to say that market price $P(Q)$ is a decreasing function of the output Q. The profit function is then given by $\pi(Q) = P(Q)Q - C(Q)$, where $C(Q)$ represents the costs related to production of electricity with capacity of Q. The profit maximising rate of production is where marginal revenue is equal to marginal cost $P(Q) + Q\frac{d}{dQ}P(Q) = \frac{d}{dQ}C(Q)$. It is typical for electricity producers to meet discontinuity in their marginal costs. In case the demand exceeds the capacity, producer can supply with a generator of lower cost, the additional capacity must be supplied by using a higher cost generator, which can be observed in the marginal cost curve as a vertical line. (Biggar et al. 2014.) In figure 5, two different scenarios of demand are presented.

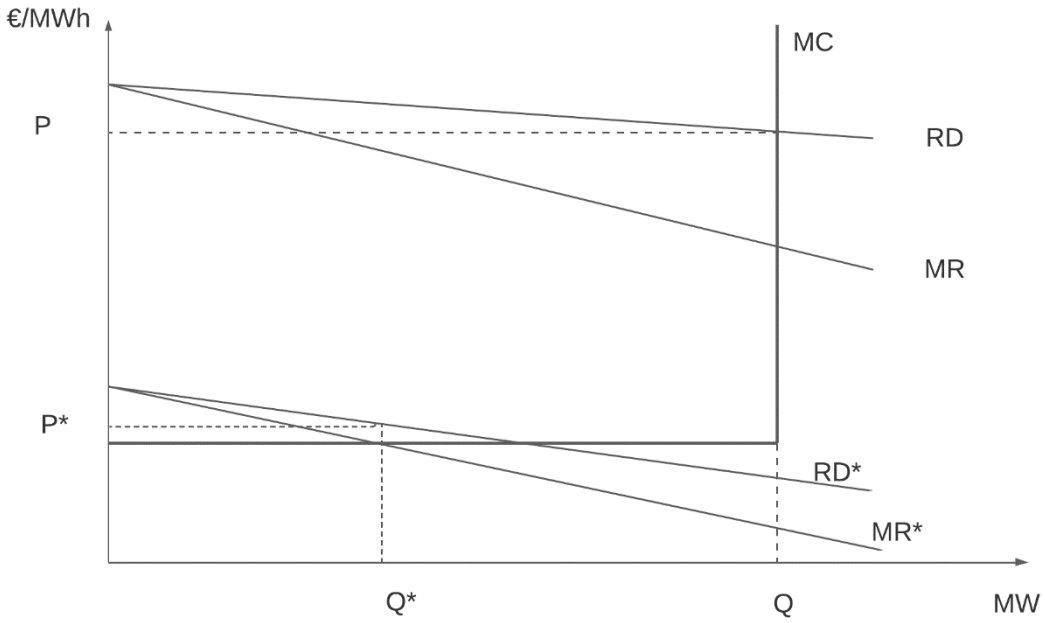

Figure 5, Exercise of market power can lead to lower level of capacity offered.

In figure 5, marginal costs (MC) are presented as a curve with discontinuity point in Q. Prices P and P* occur on quantities Q and Q* where marginal revenue (MR) intersects with marginal costs. The profit maximizing combination of the supplier is dependent on the level of residual demand (RD & RD*). In case of the higher residual demand, RD intersects the marginal cost curve at the same quantity as MR, therefore no market power is exercised, and capacity Q is sold at price P. On the other hand, when residual demand is on lower level, MR* intersects with MC at smaller quantity Q* resulting in price P*. The capacity is sold at higher price that would result from selling the capacity at level where marginal costs meet demand, the supply behaviour has direct effect on market price and thus market power is exercised. (Biggar et al 2014.)

Mansur (2008) studied social welfare in reconstructed electricity markets, which means the market design that is in use today and is described in chapter two. According to the study, short-run welfare loss can only occur in ineffective allocation of generation resources, welfare loss occurs when electricity is generated with technology that has higher marginal cost than other available technologies have. In the study, strategic market behaviour is shown to lead to welfare loss, which means, exercise of market power (Mansur 2008). Since electricity markets are prone to concentration in supply side, and exercise of market power leads to welfare loss, it is essential that markets are regulated.

In practice the suppliers do not directly choose the capacity but submit their supply curve to the market operator and get instructions on the capacity to produce. In theory, market participants can maximize their profit by submitting offers in way that causes residual demand to meet their marginal revenues. In order to do so the market participant must ensure that all its capacity does not get dispatched in the coupling. To ensure that bids are accepted only at a quantity where marginal revenue is less than marginal costs, market participant needs to limit the amount offered in the market coupling. In practice, the residual demand is unknown to all market participants, and to exercise market power, the offer curve of such market participant is set to price-quantity combinations that would lead to more production if the residual demand curve was horizontal. Two common ways of capacity withholding are economic withholding and physical withholding. Economical withholding is achieved by pricing part of the available capacity purposely sufficiently high to ensure it is left out of the coupling. Physical withholding is achieved by technically reducing the available output, for example by shutting a plant down. (Biggar et al. 2014; Crampes & Creti 2006). Another motivation to engage in capacity withholding is seeking profits via intertemporal substitution. If a market participant expects the intraday price to be higher than day-ahead price, the profits obtained for the available load can be increased by withholding capacity from day-ahead market. Intraday prices are typically higher than day-ahead. The difference between prices of the two sequential markets can be explained by risk aversion, bidding constraints, capacity constraints and market power. Regression analysis on price and quantity difference between the two markets on slope of the demand curve and dummies controlling for time has shown exercise of market power by withholding capacity from day-ahead market and selling it in real-time is shown in some price areas of Sweden (Tangerås & Mauritzen 2018). Market manipulation under supply curve competition is difficult to detect from

aggregated supply bids, and more sophisticated methods measures that include individual bidding data a market power are needed (Twomey et al. 2006). Market power of individual firm can be examined by residual demand curve cost elasticity the firm meets hourly (Wolak 2003). Market dominance on aggregate scale can be inferred from firm's share of all production, or by the relative difference of their marginal costs and bidding prices, but there is no single measure that can clearly prove dominant position of some firms in the market (Hellmer & Wårell 2009).

2.2.3 Conclusion on market power

The examples of non-genuine transactions above share a similar quality with theoretical exercise of market power. It does not matter whether the motives of a market participant lie in selling the quantity in intraday market or increasing the price for part of the offered capacity, the exercise requires market participant to engage in economical or physical capacity withholding. If there is a market price cap, market participant has no advantage in exercising physical withholding instead of economical withholding, unless legislation constrains the latter (Biggar et al. 2014). In Nord Pool there exists a price cap of 3000€/MWh, thus pointing at economical withholding. However, fifth article of REMIT prohibits uneconomical orders and transactions alongside orders and transaction placed with interest in influencing price settlement. How would exercise of capacity withholding seem in aggregated supply curve? Following formulation of supply curves as vector of price bins is authors own work and follows the idea used by Pelegatti (2013). Let P_i represent a price interval of arbitrary size and V_i all volume offered with price belonging to the interval i, where i represents whole numbers $i = (1,2, \dots, i)$. Aggregated supply curve can therefore be written as a vector of sum of bids.

$$P_i = \sum_{i=1}^{I} V_i,$$

$$ASC = (P_1, P_2, \dots, P_I)$$

The supply curves always show increasing price in relation to the volume offered. In case single market participant wants to reduce capacity sold by physical withholding, it will simply not offer all its available capacity. This means that depending on capacity withheld

by a market participant offered volume on some prices will be less than in case there were no withholding, this means that for some indexes j $V_j' < V_i$, where $j = i$. This means that

$$\sum_{i=1}^{I} ASC_i > \sum_{i=1}^{I \cap J} ASC_i + \sum_{j=1}^{J} ASC_j,$$

$$and \sum_{i=1}^{I \cap J} ASC_i + \sum_{j=1}^{J} ASC_j = P_i' = \sum_{i=1}^{I \cap J} V_i + \sum_{j=1}^{J} V_i'$$

Aggregate supply curve will show less capacity for all prices after the first price interval where capacity withholding takes place and shift downwards. The same effect to the aggregate supply curve happens if market participant does not bid all available capability by mistake. Mistakenly or purposely withheld capacity shifts the aggregated supply curve downwards. Physical withholding in situation where withholding takes place in the first price interval is illustrated in figure 6.

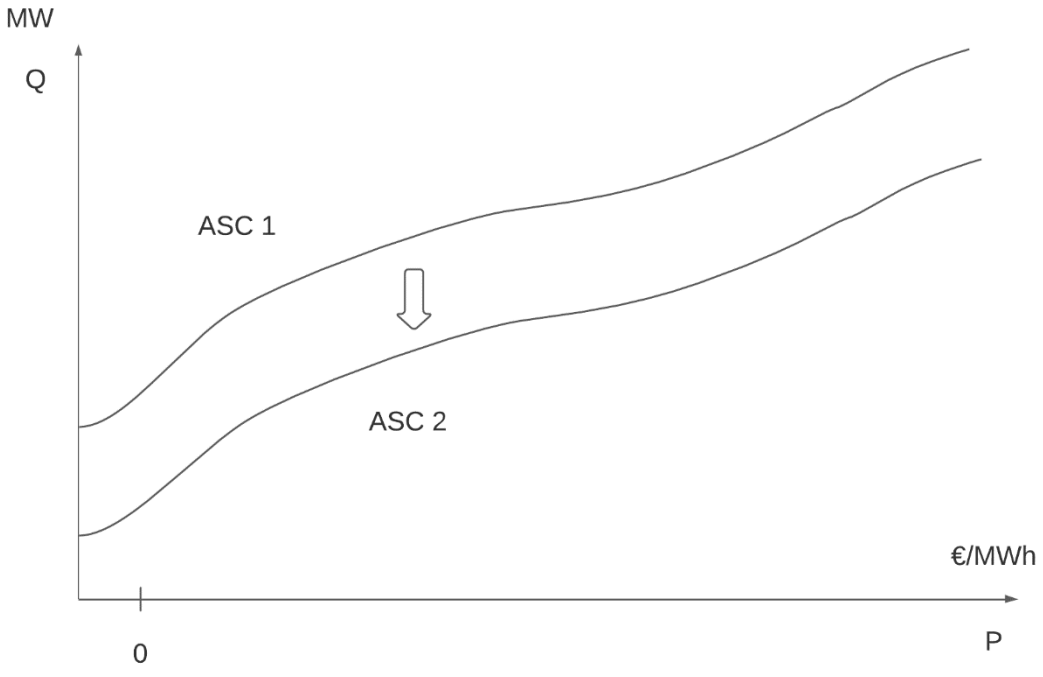

Figure 6, Mistakenly or purposely withheld capacity shifts the aggregated supply curve downwards for all prices after the one for which withholding takes place.

In economic withholding on the other hand, all the capacity available for supply is offered, but at least some of it is priced high enough for being accepted in market

coupling. This means that even though sum of offered capacity remains the same the distribution of capacity along prices is concentrated on the right side of the price axis. This is illustrated in figure 7.

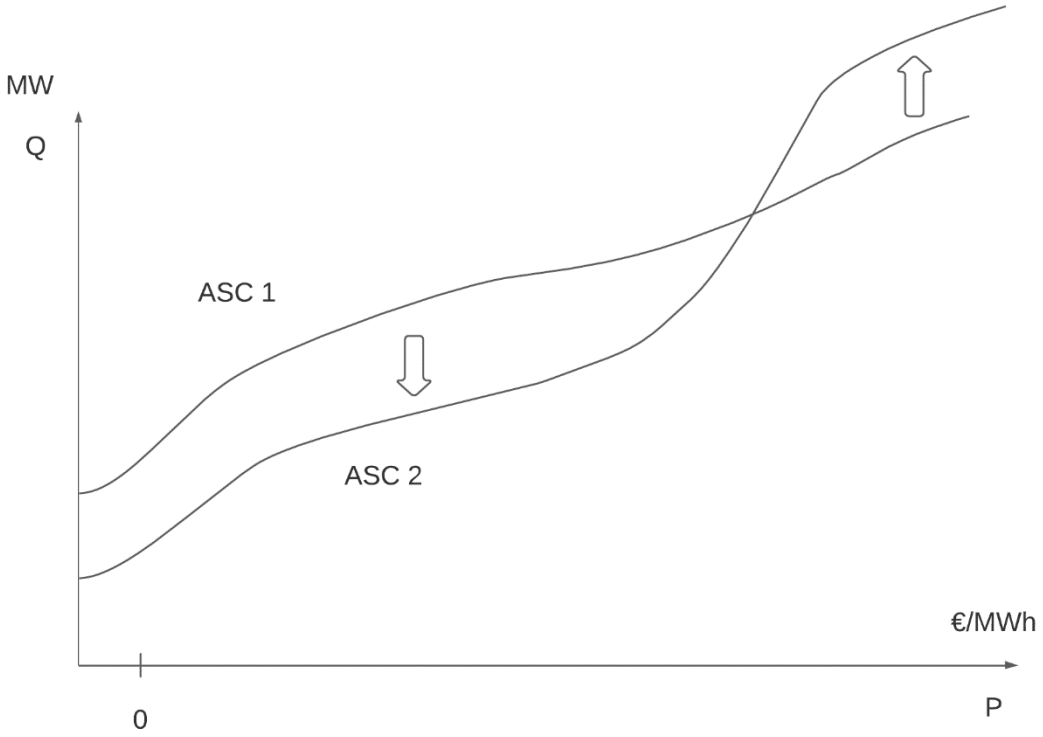

Figure 7, Economic withholding increases volume offered on higher prices

2.2.4 Market manipulation in EU legislation

Market manipulation is prohibited in European power markets by fifth article of Regulation NO. 1227/2011 of the European Parliament and of Council of 25. October 2011 on wholesale energy market integrity and transparency (REMIT). European Union Agency for the Cooperation of Energy Regulators (ACER) is tasked to enforce REMIT and further specifies that market manipulation:

> "– – includes performing false or misleading transactions, price positioning which secures or attempts to secure the price at an artificial level, as well as transactions involving fictitious devices or deception, and the dissemination of false and misleading information." (ACER 2022).

ACERs guidance on application of REMIT states that when assessing market manipulation, national regulators may identify suspicious orders or transactions that can

be held as false or misleading signals. Such transactions are further clarified to not result from genuine interest in procuring or selling a wholesale energy product at the ordered price while considering the context in which they were placed and the market participants rationale for trading. In ACER Guidance on REMIT application for regulators examples for orders and transactions that can be considered non-genuine are given, these are:

- orders placed or transactions executed at price levels that are uneconomical for the market participant;

- orders/transactions which are erroneous and therefore do not reflect a real buying or selling interest at the price considered;

- orders/transactions which are not placed/entered into with a real interest in buying or selling energy but rather with other interests (e.g. influencing behaviour of others; influencing price settlements; influencing the price of other products; circumventing market rules; benefiting positions in other contracts; tax evasion; tax fraud; profit/loss sharing; circumventing accounting rules; transferring money between market participants …);

- orders placed with no intention to execute them; and/or

- buy orders with a volume that exceeds/falls short of the buying needs/interest or sell orders with a volume that exceeds/falls short of the selling needs/interest of the market participant, in the context of its asset-backed trading portfolio. (ACER 2021.)

3 Anomaly detection with supply curves

3.1 Anomaly detection

3.1.1 Characteristics of an anomaly

To study anomalous market behaviour in unsupervised way it must be assumed that abusive behaviour does not take place most of the time. This would mean that market manipulation would not take place systematically. In case market manipulation is the norm, then anomaly detection can not bring valuable information about its exercise. If market manipulation is the norm, detected anomalies could indicate a out-of-usual situations in market fundaments or a particularly large or unusual exercise of market manipulation. However, normalized use of physical withholding would require the power plant in question to always report a failure when suitable conditions for market manipulation arises. In Nord Pool market participants are required to inform other market participants about decrease in available capacity over 100MW, it is probable that frequent downing of generation would raise suspicion. Economic withholding on the other hand is dependent on behaviour of other market participants, and frequent over-pricing of generation brings strategic risk in being excluded from coupling.

The primary assumption in anomaly detection is that the normal behaviour is stationary. Stationarity ensures that the underlying data generating process remains the same through time. If the data generating process does not change through time, it means that the characteristics that are present in the data should also be present in the future. The process behind can have seasonal, or long-term trends. (Mehrotra et al. 2017.) Anomaly can be understood as a deviation from the rule or an irregularity that cannot be considered as a part of normal system behaviour. In time series data, anomalies can be categorized to contextual, collective and point anomalies. Point anomalies, or outliers, are recognized as instances that occur on certain percentile of probability density calculated for target parameters. In other words, point-anomalies are instances that are very unlikely to result considering historical instances. Collective anomalies are group of vectors where a single vector cannot be considered to deviate from normal system behaviour, but irregular behaviour is indicated by the composition of the vector group. Contextual anomalies are individual data vectors or vector groups that are not point- or collective anomalies, but in the scope of surrounding data indicate irregular behaviour. Point- and collective

anomalies can be detected by their internal structure or content, but detection of contextual anomalies depends on accounting for the short- and long-term characteristics in the surrounding data structure. Contextual anomalies are primarily characterized by applying distance-based metrics that can be realized on sliding window technique. (Lindemann. et al 2015.)

3.1.2 Anomaly detection using the prediction error

The problem of predicting supply curve falls into category of multivariate time series prediction, where anomaly detection can be done by making inference of the prediction error. The prediction task can be formatted to supervised learning with sliding window approach (Lindemann et al. 2015). The data is formatted to have sequence of vectors as input and the consecutive vector as a target. In supervised learning terminology, targets labels mean the instance that model is trained to predict. To make inference about the prediction error, it must be assumed that all instances of the data are realizations from the same distribution. The multivariate prediction task can be described by following equation

$$f\left(x^{(t-1)}, \theta\right) = y^{(t)} + \varepsilon$$
$$E\left(y^{(t)} \middle| x^{(t-1)} = x\right) = \hat{f}\left(x^{(t-1)}, \hat{\theta}\right) = \hat{y}^{(t)} = y^{(t)} + E(\varepsilon)$$

where $f(\cdot)$ is an unknown function that describes the data generating process of $y^{(t)}$ with unknown parameters θ, ε is the error term related to the data generating process, $x^{(t-1)}$ represents D_1 dimensional input vector, $y^{(t)}$ is D_2 dimensional output vector where $D_1 > D_2$ and $(x_1^{(t)}, ..., x_{D_2}^{(t)}) = (y_1^{(t)}, ..., y_{D_2}^{(t)})$. The task is to find a function with some parameters that can approximate the function of the data generating process to estimate the target by mapping the input sequence into output in a way that can generalize also with out-of-sample data. From the conditional expectation we can see that the predicted $\hat{y}^{(t)}$ consists of the real $y^{(t)}$ and expected value of the time invariant error related to the date generation process. (Goodfellow et al. 2016, Du & Xu 2016.) In other words, the task is to estimate the function for the expected value of the target conditional on the input sequence of the data distribution. We know that prediction error should remain in the magnitude of $\hat{y}^{(t)} - y^{(t)} = E(\varepsilon)$. If prediction error gets above average, it can only be caused by a change in the data generating process, thus breaking the assumption of normal behaviour being stationary.

3.1.3 Supply curve prediction and analysis in literature

Problem in analysis of supply curves and day-ahead trading in general is the curse of dimensionality. Simple linear methods are not sufficient to generalize hourly day-ahead bid behaviour where bids are hourly and price grid for bid volume can be practically continuous.

In the literature there are several contributions on statistical analysis of day-ahead supply. Jenkin et al. (2018) use regression analysis to estimate the shape of the supply curve. Supply curves are estimated by linear and cubic fit to historical hourly realized price and load data for various tested time intervals and the analysis repeated on rolling basis. The fitted supply curves are used to analyse the effect of retirement or addition of generation. The study does not aim to predict supply curves, but to study the shape of the merit-order curve on historical data. The authors report lower predictive accuracy in hours with high loads which is speculated to be caused by curve not capturing significant changes that are due to other variables. The method's predictive capability is tested by backcasting. Two weeks of historical data is used to predict the average supply curve of the following two-week period. The cumulative average error of weighted average prices and backcast-estimated prices for entire year is reported to be 5%, it is worth noting that variance of hourly errors is mitigated as positive and negative errors cancel each other partly when the prediction is repeated for the whole year. The backcast-estimates increase variance of the results on hourly basis, and the authors recommend relying on errors mitigating in the long run. Inclusion of exogenous variables and more flexible estimation methods are recommended to improve curve fit estimates. (Jenkin et al. 2018.) In order to make inference on bidding behaviour on hourly level, regression analysis falls short. Price-load pairs provide information about the cross section of supply and demand bid curves, but the shape of the curves behind the cross section are left outside of the scope. The study answers to the question of average marginal cost curve market faces, but the method is not suitable for catching temporal trends in daily use.

Pelegatti has conducted study on predicting hourly level aggregate supply bidding curves. The bidding curves are first formed by dividing bid volumes into intervals using quantiles by price, 49 different 50-tiles are used supplemented with intervals representing minimum and theoretical maximum prices of Italian day-ahead market. The 51 intervals are then

transformed into logarithmic increments in order to force the non-decreasing shape of the supply bid curve on predictions. Two methods for predicting are presented, principal component analysis and reduced rank regression. Both methods aim for dimension reduction. In the first method based, principal component scores of log-increment price intervals are regressed on their lags and exogenous variables that include dummies for weekdays as well as sinus and cosine waves to factor in seasonality, the number of principal components that minimizes out-of-sample mean squared error (MSE) is chosen. Second method uses reduced rank regression directly on the lags of log-increments of price intervals, and similarly to the first method, choosing the rank that yields smallest out-of-sample MSE. In both methods, the predicted log-increment price intervals are applied reverse transform before calculation of the out-of-sample MSE. Both methods are tested to make 1 and 24 steps-ahead predictions using two different sets of lags as regressors. Out-of-sample mean absolute percentage errors (MAPE) are reported for all eight tested models for each day of the week. 1-step-ahead models are reported to yield MAPEs from 2.4% to 3.1% averaged over all weekdays. (Pelegatti 2013.)

Ziel & Steinert (2016) approach forecasting of day-ahead electricity price by modelling supply and demand bid curves separately and using the estimate of cross section of the two curves to predict electricity prices. In contrast to earlier approaches, bids are arranged into intervals by mean bid volume. The supply curves are modelled by high dimensional autoregressive time series model that utilizes lasso regression to estimate the coefficients. No out-of-sample prediction error is reported for predicted supply curves as the focus of the study lies in price prediction, however, the out-of-sample error that the method yields is as low as 40.6% compared to persistent model, where supply and demand of the previous week are used as forecasts. (Ziel & Steinert 2016.)

The methods for prediction of supply curves in the literature discussed vary from strictly statistical to limited use of non-parametric estimation. Another approach to estimate $f(\cdot)$ is to rely strictly on non-parametric approach, machine learning. In next chapter Neural Networks and long short-term memory network are presented. Appeal of machine neural networks in this estimation of the $f(\cdot)$ is discussed further in the chapter two.

4 Method

4.1 Long short-term memory neural network and machine learning

4.1.1 Applications of LSTM in literature

As a gated recurrent network LSTM is suitable for predicting of sequential data following the recommendation of Goodfellow et al (2014 p. 420). In literature there are multiple applications of LSTM based network for forecasting and anomaly detection of time series. Nguyen et al. (2020) suggest stacked LSTM for forecasting of multivariate time series and a LSTM autoencoder method with unsupervised learning algorithm on anomaly detection. Neural network is used to extract features from data which are then classified using one-class support vector machine which is used to subtract outliers from normal inputs. Both approaches were applied to synthetic and real data. Reported results show 12.47% mean forecast error on predicted real data, however multivariate time series are used to provide predictions for only single feature. Anomaly detection approach results for real data are only discussed qualitatively. On synthetic data the approach is reported to perform slightly better comparing to the methods in comparison, however it is only applied to univariate time series. Lindemann et al (2021) survey various LSTM network applications on multivariate and univariate anomaly detection tasks. Two main approaches on network architecture are identified: LSTM-based and encoder-decoder based. LSTM-based approach relies on stacked LSTM structure, where outputs of previous layer are used as inputs of subsequent layers and the dimension of the outputs decreases from the first layer to the last, so no dimensional reduction is utilized to extract features. The detection of anomalies is solely based on evaluation of deviation of predicted outputs by variance analysis. Encoder decoder-based methods are reported to be utilized in majority of cases in various application fields where labelled anomalies are not available. Autoencoder networks are an approach where the encoder part of the model is set to lower the dimensional representation of the data to extract representative features of the input, and decoder part is set to reconstruct the input from the decompressed features. Autoencoder framework is fit to data that represents the normal operation of the system, and therefore the reconstruction error of data including anomalies results in reconstruction error that can be used as an anomaly score. (Lindemann et al. 2021.)

4.1.2 Recurrent neural network

To understand how LSTM network works, first operation of recurrent neural network (RNN) is explained. RNN is a class of neural networks which include recurrent connections in their architecture and thus are capable of learning order dependence. RNN is specialized for processing grid of input values, like the grid of hours and price intervals. Recurrent neural networks use sequences to predict the output. Each input X_t represents a vector of p components $X_h^T = \{X_{h,1}, \dots, X_{h,p}\}$ and the hidden layer consist of K units $A_h^T = \{A_{h,1}, \dots, A_{h,K}\}$. The picture below represents the operation of recurrent neural network, where each vector in the input sequence is used to update the hidden layer. Hidden layer units use both the corresponding input vector and previous hidden layer unit as inputs, which is how RNN uses backpropagation in time to update the activation vector that is used to provide the output. The picture below represents a RNN with sequence input and single output. The multivariate forecast can be achieved by using also outputs resulting from previous hidden units, or by widening the network. RNN neuron means a unit that updates its state from the input and provides an output, such as one in the picture below. Network width means the number of RNN neurons that use the input sequence to provide a single output. The diagram at the left is a concise way to represent the operation of RNN, where the backpropagation of information is expressed as a loop. Gareth et al. 2014.)

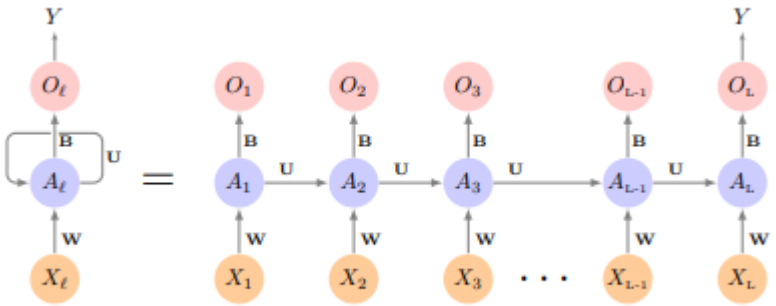

Figure 8, Recurrent neural network cell (Gareth et al. 2014 p. 422)

Figure 6 presents the recurrent structure of RNN unit. The weights **W**, **U** are $K \times (p + 1)$ and $K \times K$ sized matrices and **B** is a $(K + 1)$ size vector. They determine how each vector of the input sequence effects the hidden layer units A_t and are not functions of L, meaning that they simply determine what information from the input vectors is used to update

hidden layer units, which are also referred to as hidden states. The information of the input sequence is carried by the hidden states. (Gareth et al. 2014: 421-435.)

RNN has an inherent weakness that is known as the vanishing gradient problem, it can occur when timesteps of the model increase, it means that the gradient that is used to optimize weights of the network explodes or vanishes exponentially. RNNs with many layers have deep computational graphs where the same operation is repeatedly applied to each step of long sequence, this causes the optimization problem to contain sharp non-linearities that can prevent the weights to converge optimally (Goodfellow et al. 2016: 285). Also, the first vectors of the sequence have diminishing effect on the final output of the cell as the length of the sequence increases.

4.1.3 From RNN to LSTM

Long short-term memory (LSTM) cell was developed to deal with the issue traditional recurrent neural networks face with long-term dependencies. LSTM model is a RNN where the hidden layer units are replaced with LSTM units. Difference between LSTM and RNN units is that LSTM incorporates two tracks of hidden-layer units to compute the final hidden layer unit that is used as the activation for the output A_L. This enables the last hidden unit to receive signal from both further back in time and closer in time, so the signal from last vectors in an input sequence do not get diminished as information propagates through hidden units of the network (Gareth et al. 2014: 426).

4.1.4 Fitting a LSTM network

Fitting, or training a neural network means a process, where the outputs that the model produces for every input sequence are compared to the targets provided alongside training data and use of acquired information to update the weights of the neurons to provide outputs closer to the target. Loss function means a metric that is used to compare the output from one sequence to the target. The cost function means the average loss over the entire training dataset. Machine learning differs from pure optimization algorithms by minimizing the performance measure indirectly. In machine learning the cost function of the training dataset is minimized and the validation dataset is used to evaluate the result. The algorithm that is used to minimize the cost function is called optimizer. Optimizers are one choice that must be considered in model selection, and in addition to optimizers, learning rate of the optimizer. Learning rate quantifies the impact of the gradient update

on the weights of the network. It can be used to regularize the training process and prevent overfitting. Learning rate is crucial to model training, as lower values tend to lead directly to better generalization, but too low generalization error leads to stark increase in generalization error. In regression, mean squared error and mean absolute error are commonly used cost functions. Mean squared error puts more emphasis on bigger difference between target and output, while mean absolute error treats value of an error linearly. The fitting algorithm operates by calculating the gradient of the cost function from subsets of sequences called batches. Gradient of the cost function is calculated after each batch and is used to update the weights of the network. Batch size is one of the hyperparameters to be considered when fitting a neural network, because small batches can offer a regularization effect while larger batches provide a more accurate estimate of the gradient. (Gareth et al. 2014 p. 434-437, Goofellow et al. 2014 p. 271-275).

In order to test the performance of models with different hyperparameters, some validation data, a hold-out set, must be left aside from training for two purposes. Firstly, the validation data is used to prevent overfitting by cutting the training process as the model's loss on the validation data starts to increase. Secondly, the model's loss on validation data can be used to compare performance between different sets of hyperparameters. The loss that model yields on validation data will be referred to as validation loss. For reliable results, each model should be trained on several different training and validation sets from the same data. (Goodfellow et al. 2014 p.120) However, there is no guarantee that the data follows the same distribution through time, on the other hand, deep learning models perform better on larger datasets. Same training data can be used for validation multiple times by using different subsamples of training data as hold-out set to train the model multiple times. Walk-forward validation emulates the manner that the model is used in practice. In walk-forward validation the entire training set is divided into several different sets, and model is trained using all the historic data that has occurred before the set that is used for validation. In figure 9, Walk Forward validation with 5 folds is visualized.

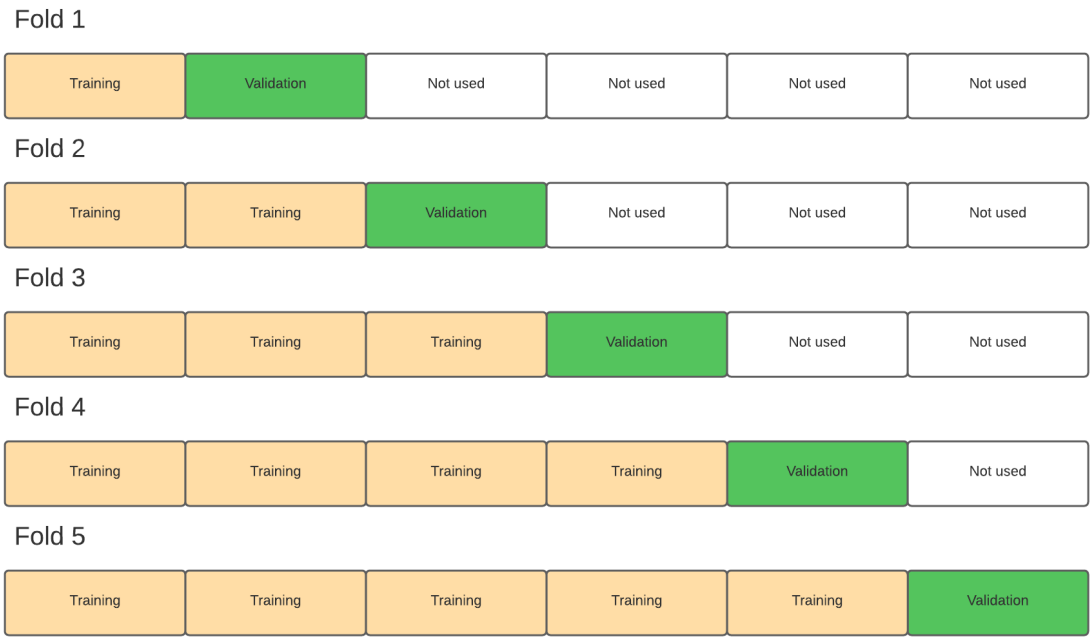

Figure 9, 5-Fold Walk Forward validation

4.1.5 Model and hyperparameter selection

There are many possible models and model configurations that can be used to estimate the output vector from input sample. In practice assumptions about the task and data have to be made to narrow down possible options. The selection of the set of models that are evaluated must rely on the previous work in the field. After the set of models is chosen, the best suitable model is chosen by model selection procedure, where different configurations are tested systematically to find the one that provides the best estimate. Goodfellow et al. recommend using train-validation-test split, where the data is split into three sets sizes of 70 %, 20 %, and 10 % respectively. Train set is used to train each model, after which their performance is tested on validation data unseen for the algorithm. After the best model and set of hyperparameters is found, training and validation data are used to train the chosen model which performance is tested on testing data. In time series environment the randomization of data before the split is not appropriate, as the systematic behaviour the model is supposed to learn is lost, and furthermore it is reasonable to use the most recent data possible for testing which reflects the nature of time series altogether. (Goodfellow et al. 2014.)

4.1.6 Anomaly detection using LSTM in literature

The function $f(\cdot)$ can be approximated via machine learning. There are numerous applications of machine learning on similar problems, ranging from anomaly detection tasks to prediction tasks. Lindemann et al. (2015) have surveyed LSTM networks on anomaly detection tasks across many fields of application ranging from medicine to pedestrian trajectory prediction and network traffic surveillance and proclaim LSTM networks to be suitable for detection of contextual anomalies due to their ability to capture non-stationary and stationary dynamics of short- and long-term dependencies. LSTM network architectures reviewed are categorized to stacked LSTMs and LSTM autoencoders as well as hybrid strategies. All three of the architectures result in some kind of error term of the predictions, which is evaluated by using either dynamic or static threshold or by application of some grouping algorithm such as support vector machine. Stacked LSTM architecture is capable of detecting collective and contextual anomalies, although LSTM autoencoder further optimize detection abilities on high dimensional data (Lindemann et al. 2015). Even though the autoencoder architecture is reported to perform better with high dimensional data, the stacked LSTM architecture has some appealing qualities. Firstly, stacked LSTM architecture is less challenging computationally, since no decoding layer has to be optimized in training. Secondly, stacked LSTM architecture has been recently implemented successfully on supply curve prediction by Guo et al. (2021). The performance of the stacked LSTM algorithm has therefore a baseline for comparison. As mentioned, stacked LSTM architecture has been utilized to predict the next instance in anomaly detection in many multivariate applications. Villarreal-Vasguez et al. (2021) use three layers of LSTM connected to a dense layer to predict next instance of very high dimensional web traffic data to classify anomalies. The sequences provided to the LSTM-model are variable in length. LSTM is used to compute the probabilities of subsequent events, which are used to classify instance as an anomaly. The target variable is categorial vector of possible events. High dimensional data is put through filtering and dimension reduction before feeding to the model. The framework reaches 95.16% rate of correctly labelled anomalies with 0.32% of instances falsely predicted to be anomalies. Tan et al. (2020) use stacked LSTM to predict anomalies from prediction error in a non-linear dynamic system. The study is conducted on univariate data representing state of a dynamic system to provide multi-timestep forecast for anomaly detection purposes.

4.1.7 Supply curve prediction with LSTM by Guo et al. (2021)

Recently Guo et al. (2021) applied LSTM neural network to predict aggregate supply curves. Motivation behind predicting aggregate supply curves comes from deriving competitions bidding strategies, in which aggregated supply curves is useful and is used as a basis for many studies on optimal bidding strategy. In fact, most of the studies about optimal bidding strategy rely on correctly forecasted aggregated supply curves. The use of LSTM for the task is motivated by successful use of LSTM for forecasting in applications across many power market areas and its qualities on dealing with relations with long lags of unknown duration. In addition to LSTM, paper introduces feature integration method to reduce enormous dimensionality of day-ahead bidding data. Study is conducted on bidding data of Midcontinental Independent System operator. Data integration is based on using price intervals that are sampled via uniform increment method, so that each price interval represents approximately same amount of volume offered. Price load pairs that do not often have effect on the market price are masked from data. Day-ahead prices in MISO realized 99.5% of the time between prices 10.14 and 77.49$/MWh, so bids with prices outside that range were not included. Dimension reduction is conducted by principal component analysis, and first 4 principal components are used as input for LSTM models. In addition to the 4 principal components, also external variables are used. External variables are called influencing factors, they are as follows: market price, load, capacity and generation mix; fuel price; season, weekday, holiday and hour of the day; temperature, irradiation and wind speed. All input data is normalized before being fed to the model. One principal component is used at a time as LSTM input in addition to external variables. Four distinct models are used to make predictions for all four principal components, which are afterwards reconstructed back to aggregated supply curves. Four distinct models are motivated by principal components not having any linear correlation by definition. Two different evaluation criteria are utilized, first for the prediction of principal components, mean squared error is used, then mean absolute percentage error is used for reconstructed aggregated supply curves. Data set comprises of 39408 hours of data and is split to training, validation and testing sets according to 80%, 10% 10% split. Hyperparameters for the LSTM were selected in grid search and every model for distinct principal components have different set of hyperparameters. With principal components the forecasts provided best results when using 48 or 72 previous hours as input, indicating that there is no relevant information to

the model beyond two days of past data. LSTM forecasts are compared to supply vector regression, random forest and multilayer perceptron with two different data pre-processing methods for each. The PCA dimension reduction and direct usage of price intervals. The performance of the models is reported in mean absolute percentage error (MAPE). All models produce testing MAPE between 2.74-6.02%, LSTM with principal components resulting the lowest. Study concludes that LSTM is very suitable for predicting aggregated supply curves, regardless of the pre-processing technique. (Guo et al 2021.) Supply curve prediction with LSTM is appealing for it can benefit from excessive dimensionality and regularize itself automatically as long as overfitting is taken into consideration.

4.2 Data and method

4.2.1 Step 1: Data and pre-processing

The data used in this study consists of price and corresponding volume of wholesale electricity supply from Finnish market cross point data, which is REMIT data collected by ACER and received from National Regulatory Agency of Finland, Energiavirasto. The dataset consists of hourly electricity wholesale price values and corresponding volume of supply offers from starting hour of 27.4.2019 00:00 to 1.11.2021 23:00 thus consisting of 22080 hours of supply curve data. The supply volumes are summed to price intervals containing the total supply volume offered for prices: from -500 € to -10 €, from -10 to 0 €, from 10 € to 100 € with increment of 10 €, from 100 € to 200 € and from 200 € to 3000 totalling up to fifteen price intervals. The leaps in the lowest and highest price intervals are used because less trading occurs on extreme prices, thus information is sufficiently represented by larger intervals.

The weather data is received from Finnish Meteorological Institute and consists of hourly time series for temperature and wind speed for the same period as supply curve data. The weather data and the volumes in price intervals are all scaled between zero and one by linear transformation.

The data used for prediction also include sin and cosine curves indicating the hour of day and day of week. Sin curves are motivated because dummy coding both day of year and hour of the year would increase dimensionality of the input unnecessarily. When both sin

and cos curves are used, each time unit gets unique value combination due to different phases of the functions. The curves are drafted from dummies representing the hour of day and the day of week by integers from 0 to 23 and 0 to 6 by using following formulas:

$$dayofweeksin_t = \sin\left(\left(2 * \pi * \frac{dayofweek_t}{7} + 1\right)/2\right)$$

$$dayofweekcos_t = \cos\left(\left(2 * \pi * \frac{dayofweek_t}{7} + 1\right)/2\right)$$

$$hourofdaysin_t = \sin\left(\left(2 * \pi * \frac{hourofday_t}{24} + 1\right)/2\right)$$

$$hourofdaycos_t = \cos\left(\left(2 * \pi * \frac{hourofday_t}{24} + 1\right)/2\right)$$

The values are thus scaled between zero and one and are well suited to be used as input of neural network. In order to format the task to supervised learning, the data is arranged into sliding windows, where the length of the window is one of the hyperparameters optimized. To hyperparameter optimization, rolling window validation is used to address the time series nature of the task.

Since there are fifteen price intervals, four time-related dummies and two exogenous variables, the data includes 21 features. There are 22080 observations. Let X_t^{P+6} represents a $T \times P + 6$ dimensional data matrix where T is the length of the dataset. P is the number of price intervals and +6 indicates the four time-related dummies, temperature and wind speed, that will be referred as the exogenous variable from now on. In order to train the neural network, the data is arranged into samples that include sequence of L prior hours of P price intervals and exogenous variables and the subsequent price interval vector, so that the input sample X_t consists of pairs $\{(X_{t-L}^{P+6}, \dots, X_t^{P+6}), X_{t+1}^{P}\}$. In this way L previous vectors of price intervals and exogenous variables are used to forecast the next price interval vector $\hat{X}_{t+1}^{P}$

4.2.2 Step 2: Model selection and validation

The Adam optimizer is chosen following the practical methodology by Goodfellow et al. (2014: 420). A common practise is to schedule learning rate to decrease along with the training, the advantage of Adam is that it adapts the learning rate automatically. However,

the initial learning rate still must be searched along with the other hyperparameters. (Kingma, D & Ba, J 2015). Stacked LSTM architecture is chosen with dropout layer for regularization and a fully connected dense output layer, and Rectified Linear Unit used as the activation. Stacked LSTM usually performs better than going for wider models (Goodfellow et. al 2014). The hyperparameters of the model are:

Learning rate,

length of the input sequence

dropout rate,

batch size,

number of units in LSTM layers.

In order to find best hyperparameters, a grid search is conducted with 5-Fold Walk Forward validation. In order to prevent overfitting, an early stopping is employed to cut the training when validation error starts to get higher while training error still decreases. Early stopping of five is used, which means that if there is no improvement in validation error in last five iterations, the training is stopped and the weights resulting in best validation error are saved. Hyperparameter space for the grid search was limited to 30 combinations of hyperparameters listed in table 1.

Table 1, Hyperparameters to be tested in grid search.

Learning rate	Dropout rate	Batch size	LSTM units
(0.005,0.0005)	(0.1,0.2)	(32,64,128)	(240,504,1000)

For each of the tested 30 combinations of hyperparameters, five different models with same set of hyperparameters are trained, one for each fold at a time. Mean absolute error is used as loss function, and for each of the five folds, validation mean absolute error and validation mean absolute error are reported. Average of MAE for all five folds of the each hyperparameter combinations are used for comparison and final selection. Also, the average of two last folds is reported and used in comparison, since the two last folds hold the largest sized training sets and might provide more realistic insight to model performance. The model with hyperparameters that performed best in the grid search is evaluated by splitting the entire data to sets with 70%, 20% and 10% of the data points to evaluate performance of the model and the generalization error. The set of 70% is used

to train the model, set of 20% is used to prevent overfitting, and the last 10% is used solely for calculation of the mean absolute error the model produces to assess the generalization error. If the model performs comparatively well, it can be used for anomaly detection.

4.2.3 Step 3: Anomaly detection

The selected model is used for anomaly detection in two ways to address collective and contextual anomalies. In order to detect anomalies via using the prediction error as measurement, one must assume that validation and training data come from the same distribution. If the distribution remains the same in the data yet unseen for the model, prediction errors should remain under a threshold of average error that can be deducted from the validation data. As the anomalies are not labelled beforehand, and there is no information if there are anomalies in the data, it is reasonable to apply walk forward method to see how prediction error behaves when model gets more data to use for training. When model is trained in walk forward method, mean absolute error should remain averagely the same for validation set and data that has not been used for training or validation. If the model is proven to generalize well and prediction error is higher than average for unused data, it can be interpreted to result from different data distribution and thus be labelled as anomaly. This way the prediction error can be inferred to answer to the question, is the data anomalous conditional on the historical data to that point in time. The fourth fold can be used to evaluate the generalization error of the model and compare it to the LSTM prediction method used by Guo et al. (2021), since it uses 4/6 of the data to training, 1/6 of data for validation and 1/6 is left and can be used for out of sample testing. The anomalies are labelled with visual inspection of the error curve and on basis of mean absolute error they produce during testing. If mean absolute error for single hour exceeds average for validation set, it is inspected as anomaly.

Threshold setting strategy in anomaly detection usually relies on two strategies. Predefined or posteriori. It is difficult to define reliable predefined threshold for normal condition of multivariate time series. Usually, posterior threshold is defined by inspecting receiver operating characteristics curve (ROC). ROC curve is a scatterplot of true positive rate and false positive rate and thus cannot be used in the absence of prior knowledge of labelled anomalies. Besides, for anomaly detection, confusion matrix of false positive, false negative, true positive and false positive classifications leads to more reliable results.

(Liang et al. 2021). In a setting where there is no previous information about the distribution of anomalies in the dataset one can rely on Gaussian assumptions about distributions of past smoothed errors. Computationally costly alternative is to use machine learning model can be trained to compare errors, such as k-nearest neighbours or support vector machine. Using unsupervised machine learning however reduces the interpretability that is already low when using black-box models for prediction.

5 Fitting and estimation

5.1 Model selection

5.1.1 Steps 1 and 2

The hyperparameter search was conducted for 32 different combinations of hyperparameters. Early stopping of 5 was used with Adam optimizer. L of the input is set to 24, so previous data vectors of one day are used as input for the consequent supply curve. Model consists of two LSTM layers and the architecture was set up as discussed earlier. The combinations that resulted in best validation error on average of all five folds are listed in table 2.

Table 2, Average validation loss for hyperparameters of all five folds from hyperparameter search.

Hyperparameters	Validation MAE	Validation MSE
[1000, 0.2, 32, 0.0005]	0.012908	0.000284
[1000, 0.1, 32, 0.0005]	0.015894	0.000449
[1000, 0.1, 64, 0.0005]	0.016688	0.000494
[504, 0.1, 32, 0.0005]	0.017051	0.000524

Hyperparameters reported in the represent number of LSTM units, dropout rate, batch size and learning rate respectively. Number of LSTM units represents the first LSTM layer, the second LSTM layer has been set to include exactly half the units the first layer has. Wider models, which are models with higher number of LSTM units, seem to have outperformed narrower models. The smaller learning rate of 0.0005 is present in all best combinations, indicating that smaller updates to the weights help the training process to optimize the loss function. To get better understanding on how the hyperparameters qualified with larger datasets, specifically with folds number four and five, the hyperparameters with best averages of the last two folds where 80% of the data and all of the data are used for training and validation are reported in table 3.

Table 3, Average validation loss for hyperparameters of all five folds from hyperparameter search.

Hyperparameters	Validation MAE	Validation MSE
[240, 0.1, 32, 0.0005]	0.012103	0.000252
[504, 0.2, 64, 0.0005]	0.012657	0.000276
[1000, 0.1, 64, 0.0005]	0.012681	0.000277
[1000, 0.1, 32, 0.0005]	0.012715	0.00028

When only the average of two folds where the model is trained with all or almost all of the data are compared, it seems that narrower models outperform wider models. The narrower models might learn the task better with bigger datasets because the model weights need to adjust more carefully to generalize well, wider models come with more modelling power and are capable of adjusting to the training set faster, which also means that the danger of overfitting the model is more potential with wider models. Early stopping algorithm should overrule the possibility of overfitting, but closer inspection on learning curves of the two models is required to assess their suitability. In figure 10, the MAE loss of the models with best averages is plotted for training and validation data for each training epoch.

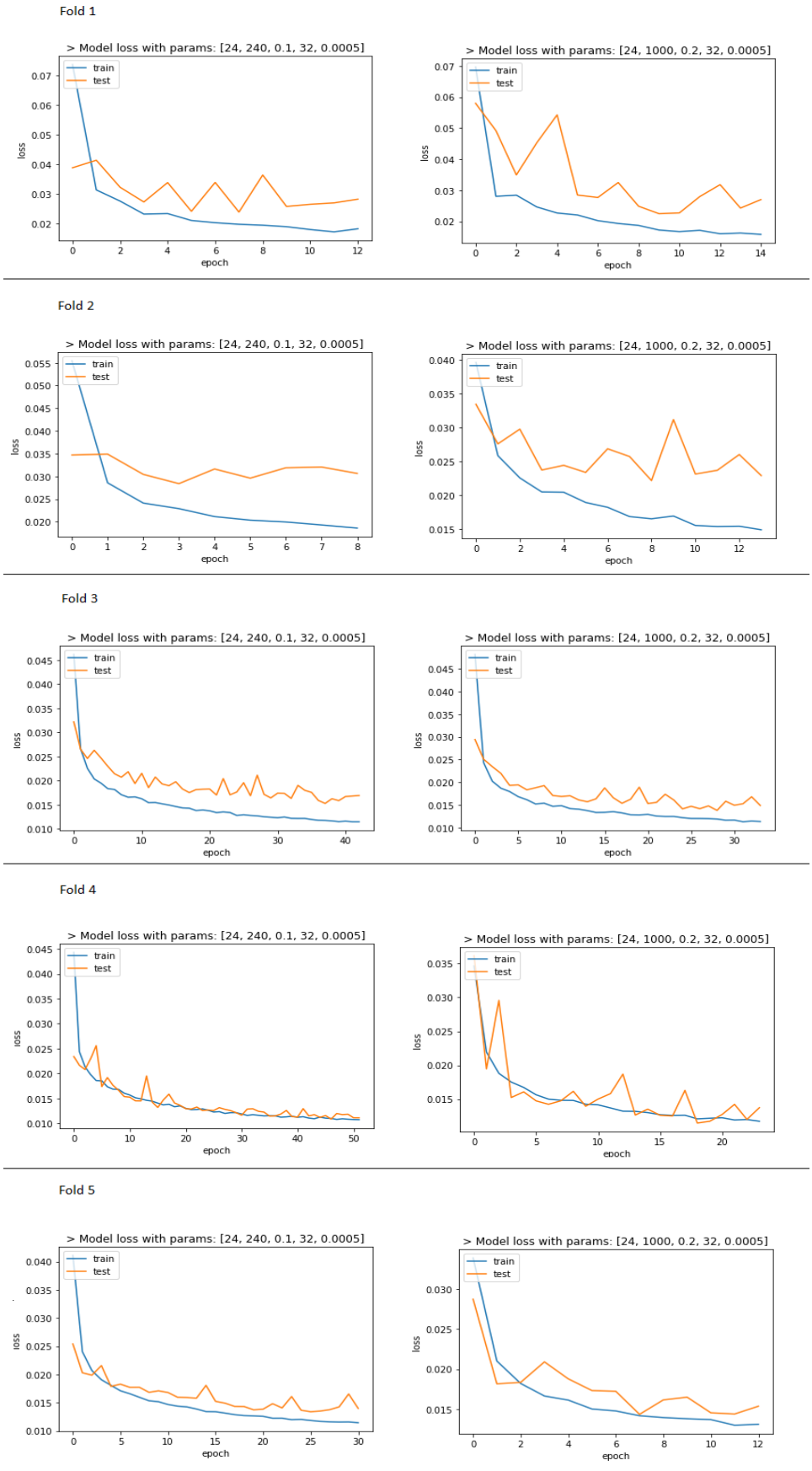

Figure 10, Learning curves of the models with best performing hyperparameters for all 5 training folds.

In figure 10, both models seem to converge to the lowest validation loss quickly in the folds with lesser data. Narrow model shows much steadier training across all folds, but the validation loss it provides for the bigger folds is only slightly smaller than for the wider model. Wider model is computationally more demanding but seems to converge to the minimum validation loss in less epochs than the narrow model. The wider model with hyperparameters [1000, 0.2, 32, 0.0005] is chosen, and used for anomaly detection. For every hour, anomaly score is calculated. MAE of each supply curve is reported as h-MAE, it is the sum of absolute values of prediction error for each price interval divided by total number of intervals which is 15. It is worth noticing that average of h-MAEs over the set used for validation is equal by definition to the validation loss.

$$h - MAE_t = \sum_{i=1}^{15} |\hat{X}_t^i - X_t^i| \,/15$$

5.2 Evaluation of the model and data distribution

5.2.1 Fold 1

Model is first trained using 3676 first hours for training and hours from 3677 to 7352 for validation. Total number of epochs was 11 and training time was around 40 minutes with 2s/step and 115 steps per epoch. Validation loss is 0.0252 which means the average over h-MAEs of validation set. Average anomaly score for the unseen data, hours between 7353 and 22057, is slightly higher, 5248. This indicates that the data unseen for the model does not follow the same distribution or that the model is not capable of capturing all non-linear relations in the data. From figure it is easy to identify that distribution of h-MAE is increased after 10000 hours and falls down approximately two months later. This part shall be referred as possible collective anomaly for now on. There are also some notable spikes around 15000 hours. The statistics of the first fold in table 4.

Table 4, Model training statistics of fold 1.

Epochs	Training loss (MAE)	Validation loss (MAE)	Training MAPE	Validation MAPE
6	0.0190	0.0252	3.53	3.84

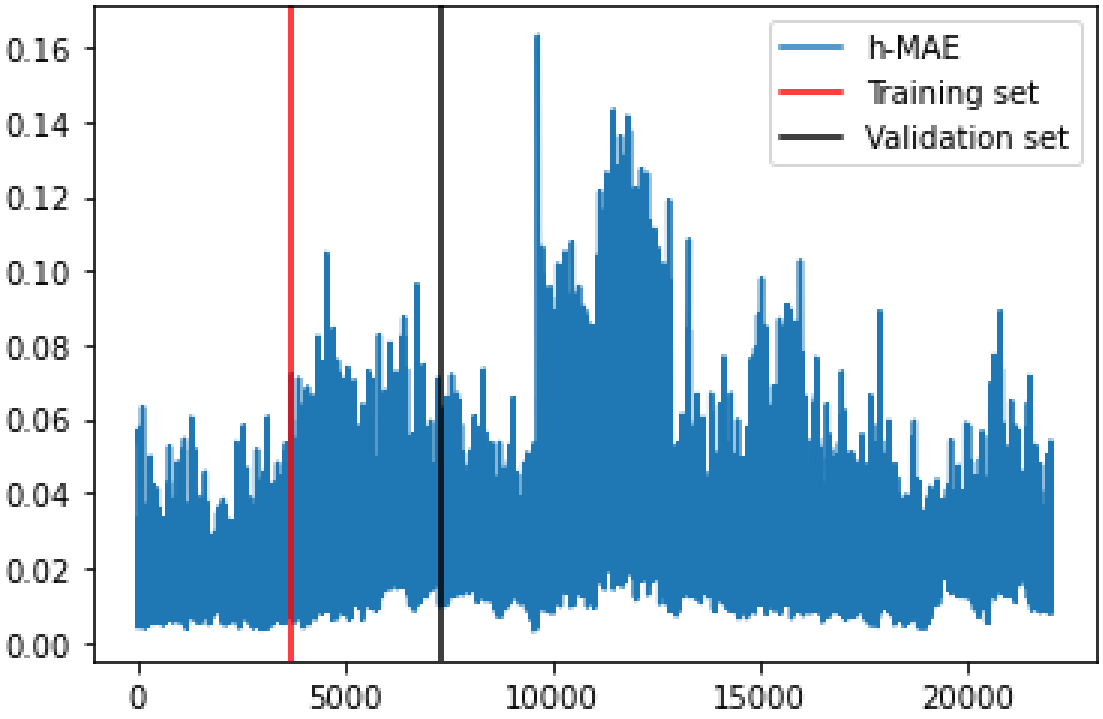

Figure 11, Timeseries of h-MAE in the first fold of walk forward validation. Vertical lines indicating the index of last hour in training and validation sets.

5.2.2 Fold 2

Now model is trained using first 7352 hours and validated on hours between 7352 and 11028. Total number of epochs was 18 and training time around 114 minutes with 2s/step and 230 steps per epoch. Now average of h-MAEs outside training and validation set is 0.204, which is almost the same as for the validation set, thus indicating that model can approximate distribution of supply curves conditional to previous supply curves well. Now part of the possible collective anomaly was part of the validation set, and model seems to have learned to expect it partly, since h-MAE significantly increases in the beginning of possible collective anomaly and again after last hour of validation set. Statistics of the round in table 5.

Table 5, Model training statistics of fold 2.

Epochs	Training loss (MAE)	Validation loss (MAE)	Training MAPE	Validation MAPE
14	0.0147	0.0203	2.5401	5.0044

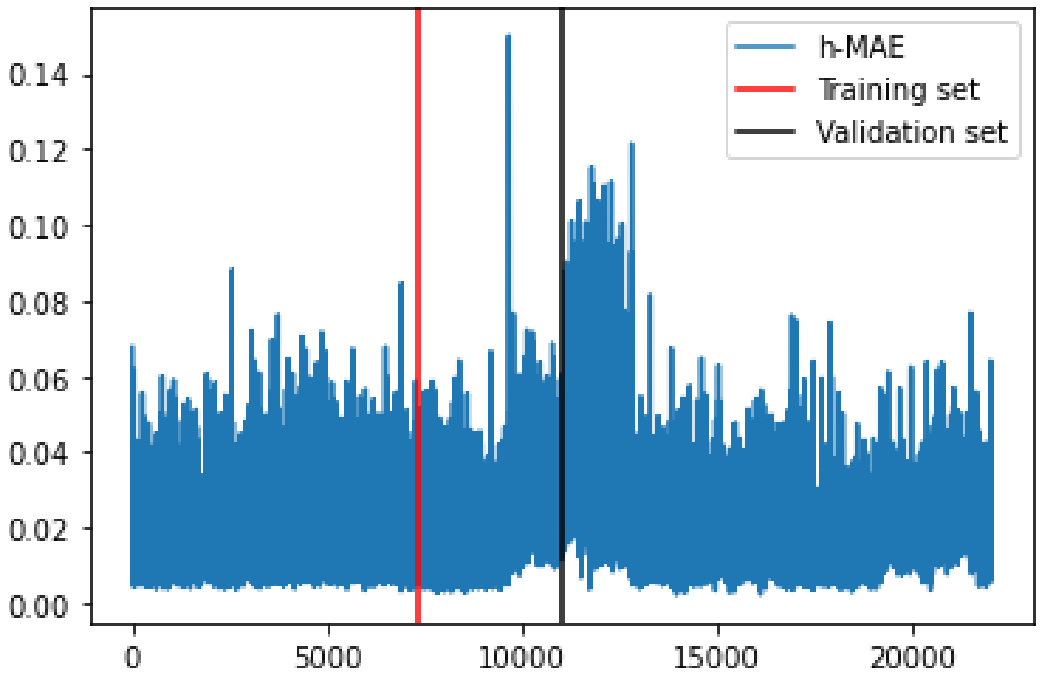

Figure 12, Timeseries of h-MAE in the second fold of walk forward validation. Vertical lines indicating the index of last hour in training and validation sets.

5.2.3 Fold 3

Now first 11028 hours were used for training and hours between 11028 and 14704 for validation. Average of h-mean for unseen data is now 0.0134, which is slightly less than validation loss. Now half of the possible contextual anomaly was included to the training and model clearly adapts to it well, still producing big h-MAE for the first hour of its occurrence. Total number of epochs was 28 and training time around 322 minutes with 2s/step and 345 steps per epoch. Statistics of fold 3 in table 6.

Table 6, Model training statistics of fold 3.

Epochs	Training loss (MAE)	Validation loss (MAE)	Training MAPE	Validation MAPE
28	0.0126	0.0146	2.3335	3.6605

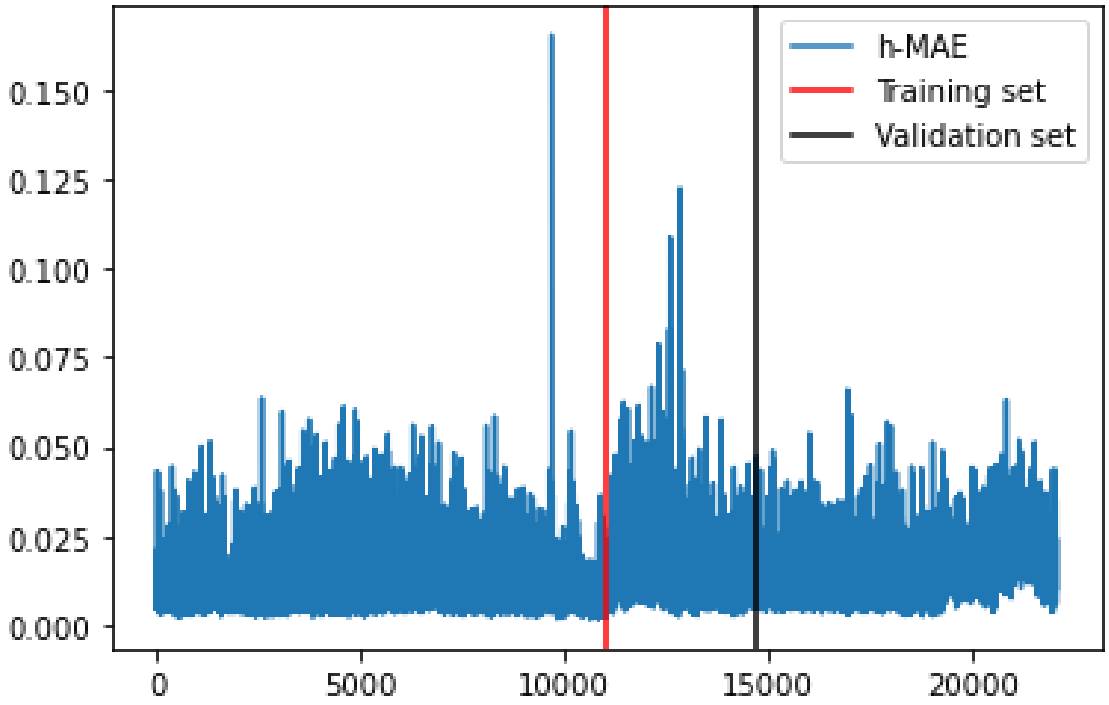

Figure 13, Timeseries of h-MAE in the third fold of walk forward validation. Vertical lines indicating the index of last hour in training and validation sets.

5.2.4 Fold 4

In fold 4 first 14706 hours were used for training and half of remaining hours is used for validation. Average h-MAE of unseen data is 0.0144, which is now higher than for the validation set. Interesting in fold 4 is that training loss is less than validation error, which can indicate that there were difficult samples in training compared to the validation set. This could result from the possible contextual anomaly, which now is included wholly in the training and validation sets. Total number of epochs was 20 and training time around 306 minutes with 2s/step and 460 steps per epoch. Statistics of the fourth fold in table 7.

Table 7, Model training statistivs of fold 4.

Epochs	Training loss (MAE)	Validation loss (MAE)	Training MAPE	Validation MAPE
15	0.0129	0.0119	2.5351	2.4092

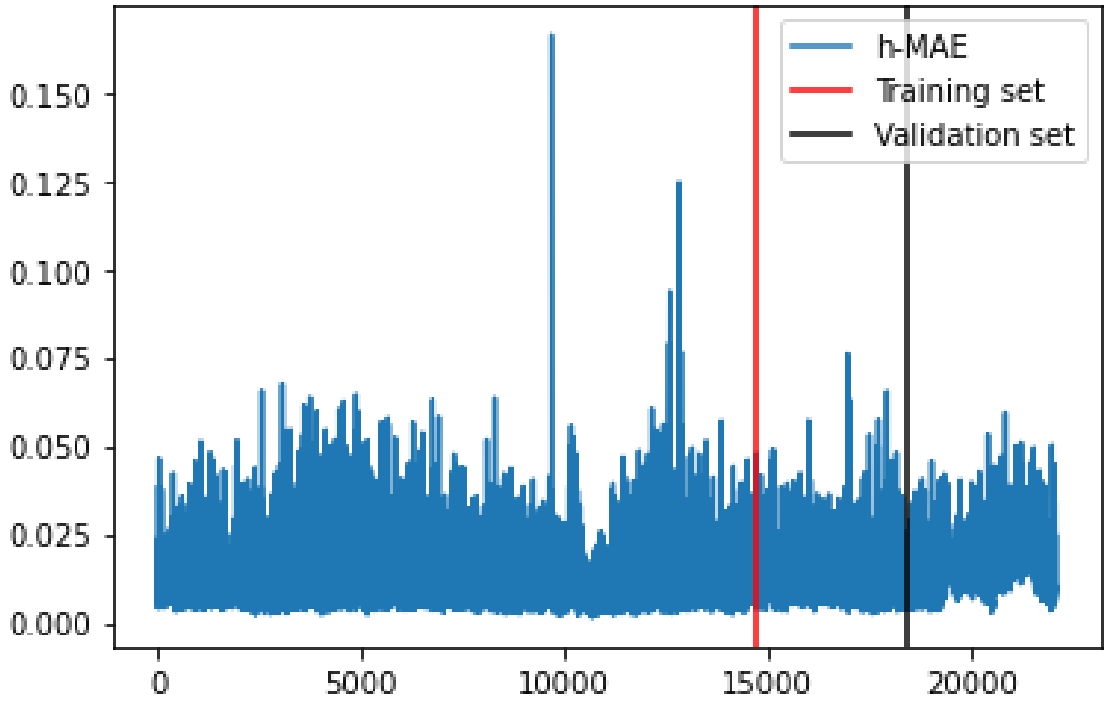

Figure 14, Timeseries of h-MAE in the fourth fold of walk forward validation. Vertical lines indicating the index of last hour in training and validation sets.

5.2.5 Fold 5

In the final round there is no unseen data to assess generalization of the model with. This is the model that will be used for anomaly detection relying on results about generalization error the previous folds provided. All expect last 3676 hours are used for training and the rest for validation. Model produces average h-MAE of 0.0107 over entire dataset.

Total number of epochs was 23 and training time around 440 minutes with 2s/step and 575 steps per epoch. Statistics are reported in table 8.

Table 8, Model training statistics of fold 5.

Epochs	Training loss (MAE)	Validation loss (MAE)	Training MAPE	Validation MAPE
18	0.0120	0.0136	2.2857	2.8358

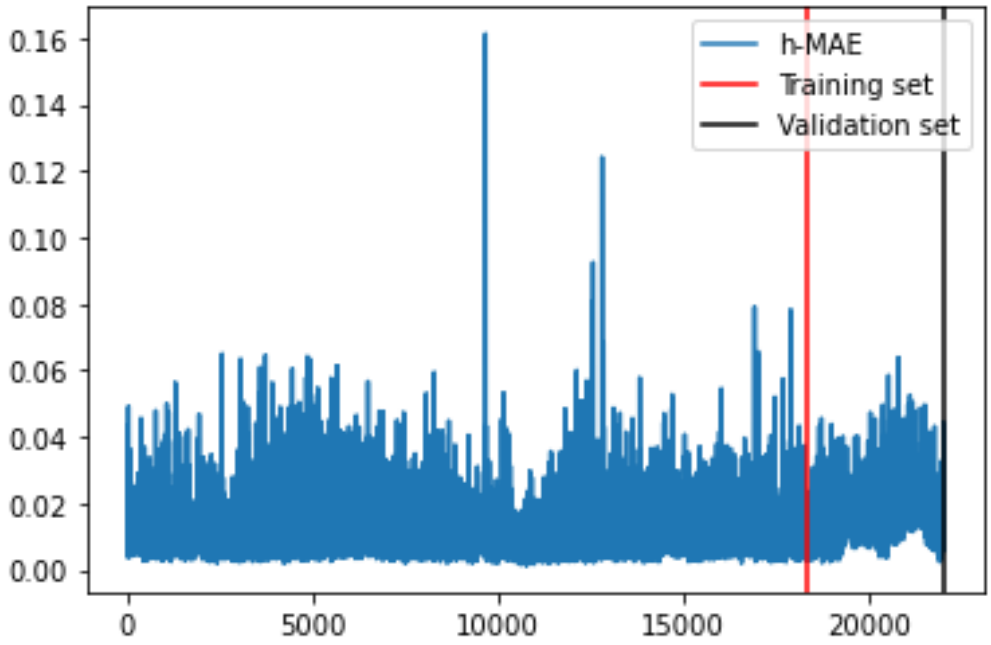

Figure 15, Timeseries of h-MAE in the last fold of walk forward validation. Vertical lines indicating the index of last hour in training and validation sets.

6 Results and discussion

6.1 Examination of distribution of h-MAE

6.1.1 Examination of hourly prediction errors

The predictions seem to show some systematic seasonality in error. In the figure 16, average h-MAE is calculated for every hour of the day over the entire dataset, in the horizontal axis hours are presented by 0 meaning 00:00-01:00, 1 meaning 01:00-02:00 and so forth.

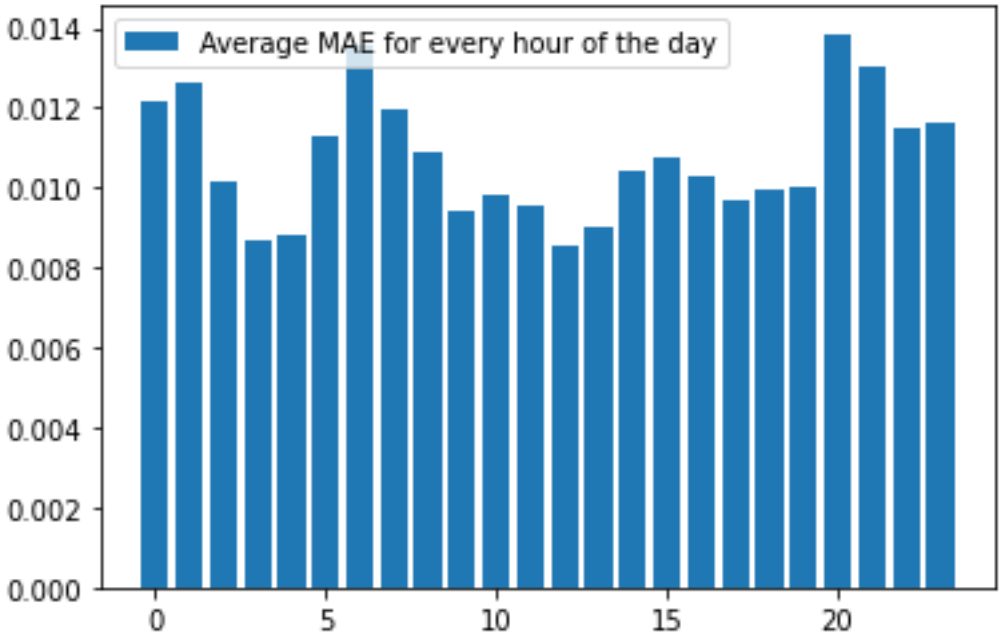

Figure 16, Average h-MAE according to each hour of the day in data.

Hours 01:00, 06:00 and 20:00 stand out with bigger average error. These hours have been harder to predict in average. Systematic seasonality in predictions imply that daily patterns present stationary behaviour the model has not been able to capture. In order to inspect prediction errors, the 25 hours with biggest h-MAE will be highlighted as a point of interest. There are 25 hours that have h-MAE higher than 0.5789. In figure 17 the 25 hours with biggest h-MAE are highlighted in red from the time series of the h-MAE.

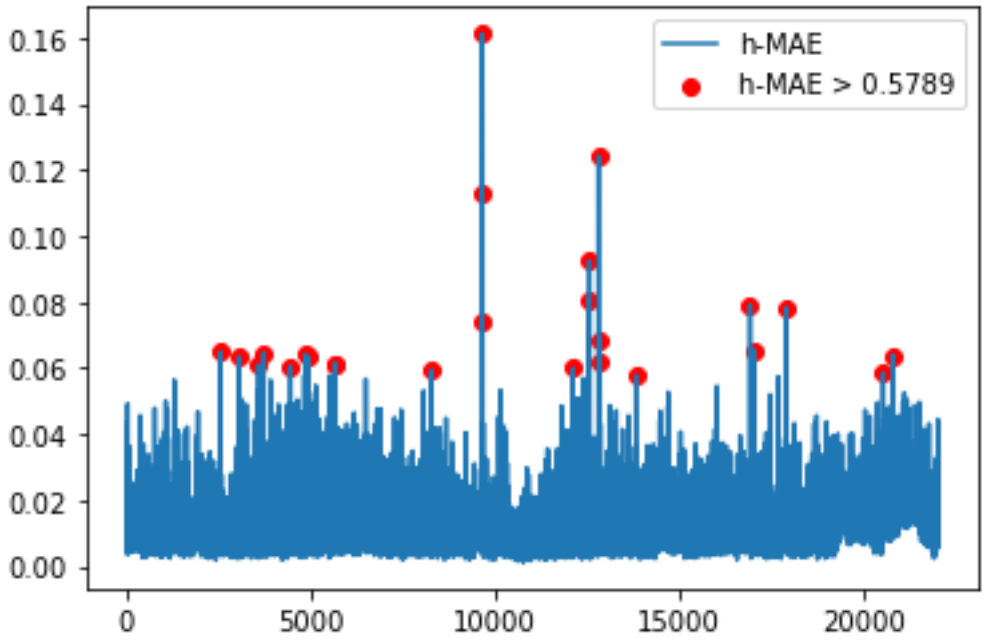

Figure 17, Time series of h-MAE and 25 hours with highest h-MAE highlighted.

The hours around the possible collective anomaly are affiliated with considerably high h-MAE in figure 17. Marked hours from 3062 to 5675 show a pattern of high errors, but before them comes a peak of h-MAE at 2567, in a period of otherwise low h-MAE. Also, the two peaks at 16934 and 17909 indicate something unsuspected considering the context of otherwise low h-MAE. The histogram of h-MAE show that the distribution of h-MAE is highly skewed to the right, and most of the density concentrated around the mean of 0.0107. In figure 18 the histogram of h-MAE, and a red line indicating the 25 hours with highest error.

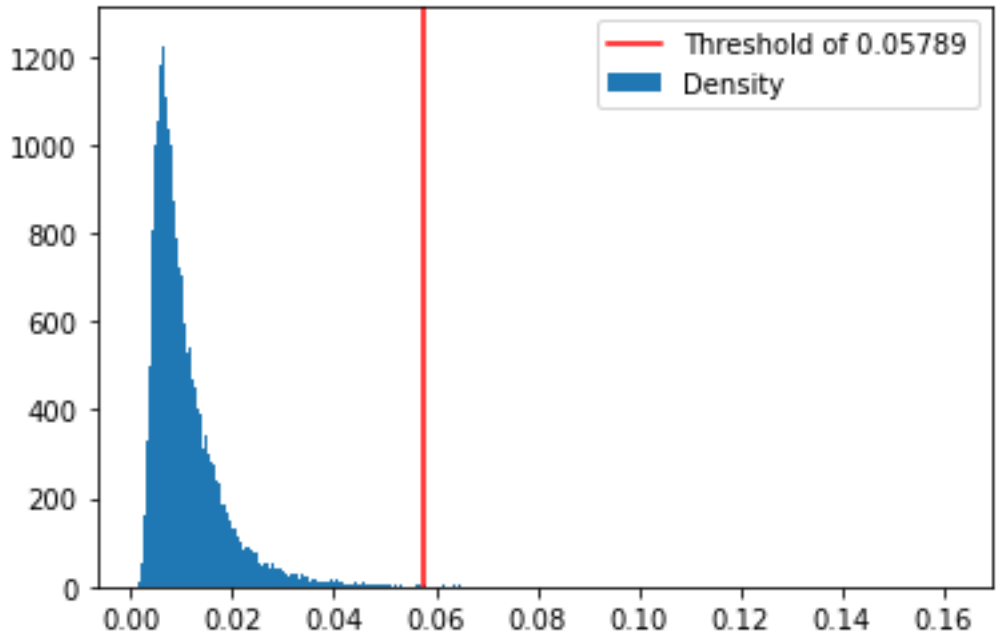

Figure 18 Histogram of h-MAE with threshold of 0.05789 indicating 25 hours with highest h-MAE.

From the histogram in figure 18 it can be seen that due to high skewedness of the distribution it is difficult to apply a clear threshold to label hours anomalous on the basis of h-MAE alone. However, the skewedness of the h-MAE implies that since larger errors are not distributed evenly, the model captures some stationary behaviour and thus larger prediction errors result from deviation in the data.

6.1.2 Examination of hours with 25 highest h-MAE

The 25 hours with highest h-MAE can be categorized to three types. First type includes the hours where predicted curve follows the shape of the true curve but is shifted vertically to lower level of volume for all price intervals. Second type are the hours where predicted curve is shifted to higher level of volume for all price intervals. Third type are the hours where shape of the predicted curve does not follow the shape of the true curve. The marked hours from 3052 to 5675 excluding the hour 4460 represent the type 1 and share the same pattern of predictions systematically underestimating the volume for each price interval by around 800 MW. In figure 20 hour 3062. Also, the hours from 12817 to 12861 and 13834, 16934 and 17030 are type 1 with varying shift downwards. In figure 19, typical pattern to type 1 marked hours is observable in hour 3062. Predicted curve in blue follows the shape of the true curve in orange on a lower level of volume. Difference of true and predicted volumes in green.

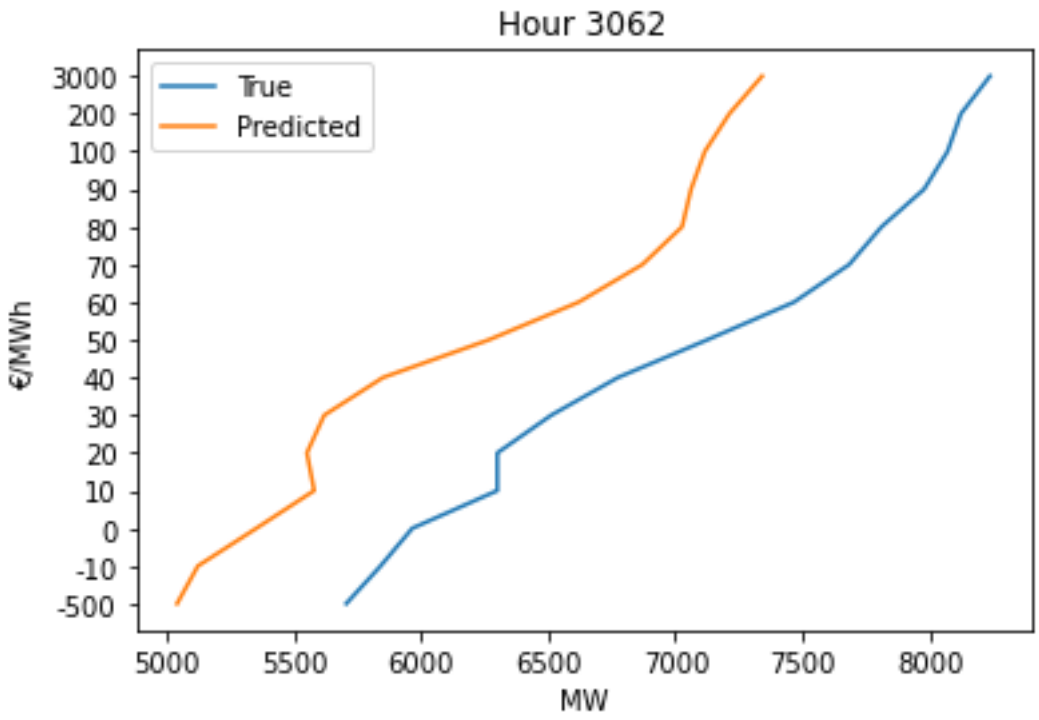

Figure 19, Hour 3062 represents type 1 of the categorized highest errors with predicted curve shifted downwards from the true curve.

There were 13 type one hours out of the 25 highest prediction errors. Next come the type 2 errors, where prediction overestimates the true volume, but the shapes of both curves are similar. Type 2 hours are 8488, hours from 9673 to 9675, 12577, 12817, 17909, 12534 and 20816. The last category is type 3, which includes hours 2567, 4460, 4930 and 12135. Hour 2567 shows a sharp increase in volume of offers priced between 20€ and 30€. Hour 2567 in figure 20.

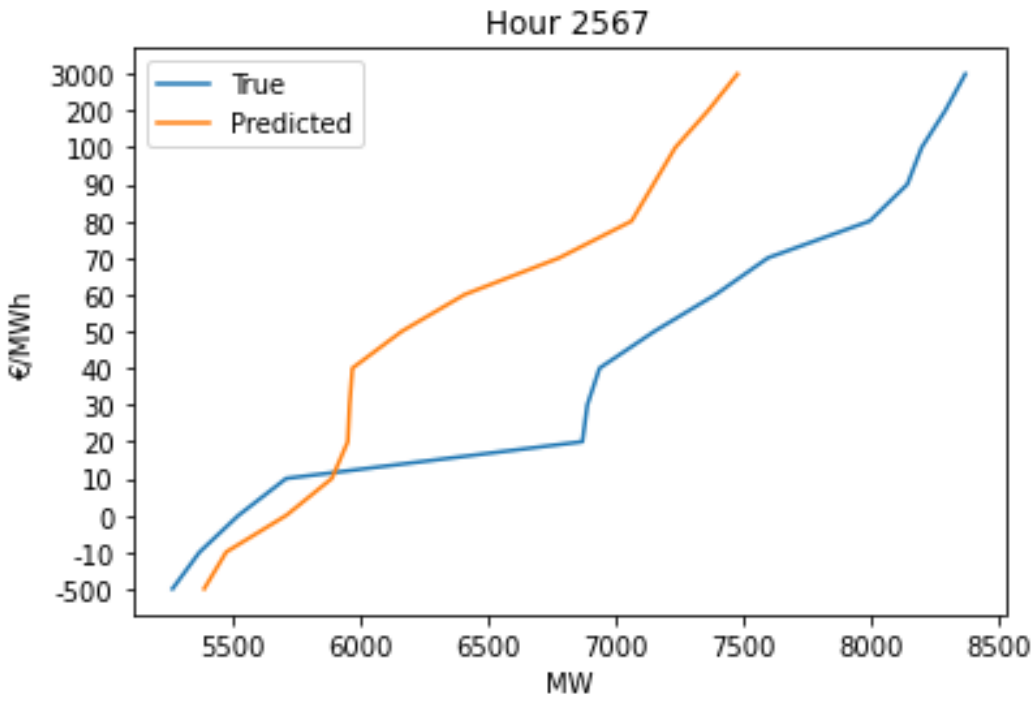

Figure 20, Prediction error of the hour 23:00 on 12.8.2019

Clear jump in volume is visible in figure 20. The rest of the type 3 hours show pattern with prediction error slightly increasing or decreasing along the price interval. One of them is plotted in figure 21.

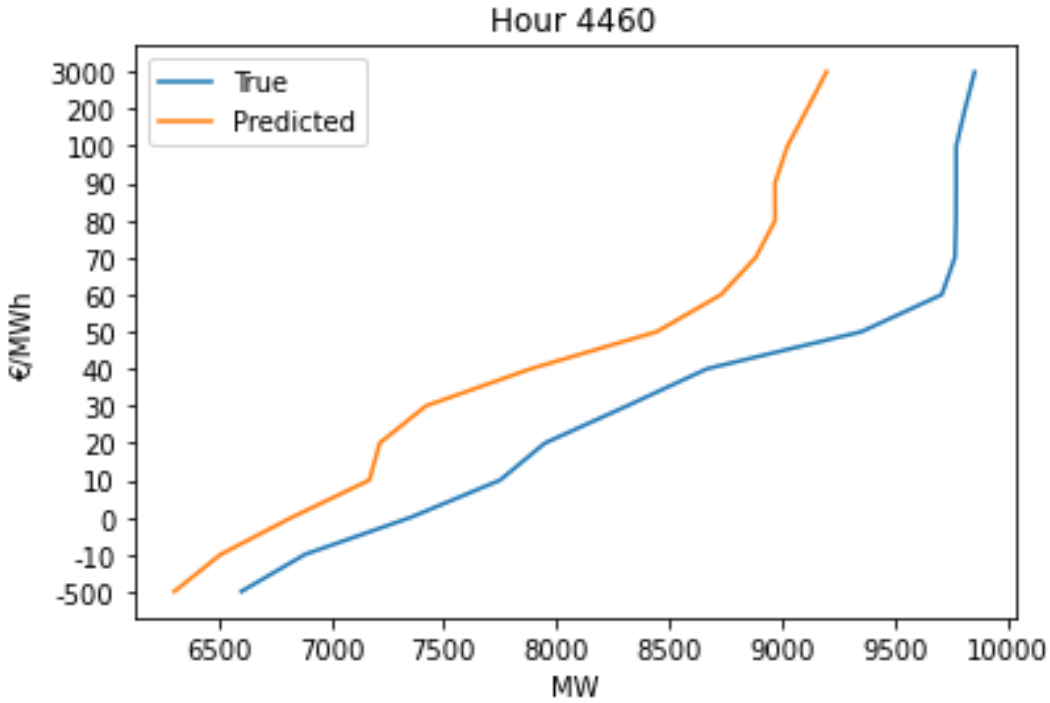

Figure 21, True and predicted supply curves differ in position and shape on hour 4460.

The three different cases among the hours with highest h-MAE share some qualities that can be further studied taking the entire dataset into consideration. Firstly, in type 1 and 2 hours there are no intersections between predicted and true supply curves and the difference between the curves along the price intervals remains approximately fixed. In type three there can be one or more intersections, but the difference between predicted and true curves is not fixed along the price intervals.

6.1.3 Examination of the three types of hours.

In the examination of hours that associated with 25 highest h-MAE three categories became apparent. In order to understand how the model performs, a further analysis on the prediction errors is provided by incorporating analysis of the three identified categories to the entire dataset. In chapter 2 possible causes of anomalies were identified as physical withholding, economical withholding, erroneous orders and data errors. To quantify the relation between predicted and true supply curves, all hours in the data are categorized according to how many times the two curves intercept. Analysis of the interceptions of two curves is also motivated by theoretical concepts in exercise of market power. In economical withholding the mass of the true curve should be shifted towards higher prices, which should mean that predicted supply curve intersects the true supply curve at least once. In table 9 the hours are categorized based on intersection points between predicted and true supply curves.

Table 9, Hours categorized according to on how many points predicted and true supply curves intersect.

Intersections	0	1	2	3	4	5	6	7	More
Hours in category	4933	3080	5465	3937	2443	1369	578	168	84

The majority of the hours have less than two intersections. It is to be expected that the average h-MAE is lower among the hours with more intersections. In figure 22 boxplots are drawn for h-MAE of hours according to their intersections.

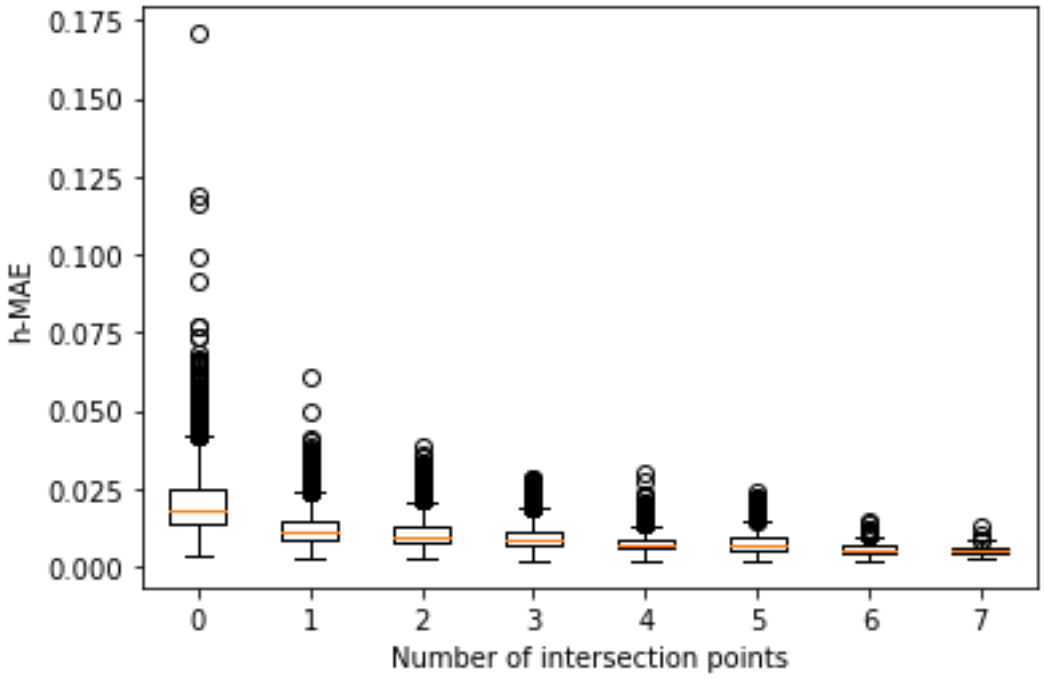

Figure 22, Whisker boxes of distribution of h-MAE on hours categorized by the number of intersection points between true and predicted supply curve.

From the boxplots in the figure 22 it is apparent that the largest values of h-MAE are strongly represented in the hours where there is one or no intersection between predicted and true supply curve. Highest values of h-MAE appear in curves where there is no intersection at all. Although it is logical that high prediction errors result from curves that do not intersect, the lack of intersections could also indicate capacity withholding or surprising events that reduce or increase offered capacity. Type one and two hours previously discussed can be studied further by observing the variance of the division of the predicted and true volume for each price interval. If the shape of the predicted curve follows the true curve, variance of the vector that results from division of true and predicted volumes for each price interval remains proportionally low. In other words, if the variance of prediction errors of one hour is relatively low. In table x, the number of hours with variance below quantiles of variance of prediction errors in the entire dataset is reported for hours that do not have intersections, which there were 4933 in total.

Table 10, Variances of division of true and predicted curves of hours with no intersections.

Quantile	**25%**	**50%**	**75%**
Variance of the hour's prediction errors	5739	9071	14701
Number of hours with no intersections in the quantile	1853	3210	4255
Share from all hours with no intersection	37.6%	65.1%	86.2%

In table 10 distribution of the errors type one and two are those whose prediction error's variance is below first quantile. There were total of 1853 hours of type one and two. Mean h-MAE in type one and two hours is 0.015 while for entire dataset mean h-MAE is 0.012. Maximum h-MAE of the type one and two hours was 0.115 occurred on hour 12817. In the table, possible type 3 hours belong to hours whose variance is below median or third quantile. Type three hours are also those that have one intersection point between predicted and true curves. Mean of variance of hour's prediction errors in hours that have one intersection point was 0.012, which equates the mean variance of hour's prediction errors in the entire dataset.

6.2 Qualitative survey over hours with highest errors

6.2.1 Hours with the highest errors

The hours with high h-MAE were divided evenly along the data, with few visible peaks. To further examine factors that resulted in low predictability, more enquiries need to be made about bidding of individual firms. There were only few public verdicts about REMIT violations in Finland between 2019-2021. Possible explanation for those hours where offered volume was overestimated can be outage of power plants or transmission capacity.

6.2.2 Hour 2567

Hour 2567 is the hour 23:00-00:00 on 12.08.2019. Energy authority of Finland has issued a verdict about trading of RAO Nordic on the same day. RAO Nordic mistakenly issued

a bid of 1060MW for the hour 23:00-24:00 CET while other bids for the same day were either 0MW or 140MW. In addition to the 1060MW mistakenly offered to the day-ahead market, RAO Nordic also sold 1060MW via bilateral contract on that specific hour. It was not technically possible for the company to supply total of 2320MW, due to available transmission capacity from Russia and the fact that the RAO Nordic did not possess any production of its own. All the capacity offered to the day-ahead market was accepted in the following day. (Energiavirasto 2021). This mistakenly bid excess 1060MW is clearly observable in the prediction error of figure 21. Prediction error sharply increases by 1100MW at price interval of 20-30€. The total h-MAE model results for the hour in question is 0.06, which is the 23rd largest prediction error in the entire dataset.

6.2.3 Hours between 9673 and 12817

Hours between 9673 and 12817 are summer hours starting from 4.6.2020 00:00 and ending in 13.10.2020 00:00. This period was difficult to predict out-of-sample in folds 1, 2 and 3. Also its beginning and ending raise high prediction errors in-sample in folds 4 and 5. The period was speculated to be a possible collective anomaly. Further inquiries about the period show that the volume traded in each price interval drops around 2000 MW compared to the time outside the period. The drop of 2000MW equals the size net value of imported energy. Further inspections to the data behind price intervals show that the net values of imported energy are completely missing during that period, which indicates to an error in data, since there was no large-scale interruption in exporting or importing electricity during summer 2020.

6.3 Discussion

Search of hours that indicate market manipulation or data errors remains difficult. Although means of market manipulation as exercise of market power are distinguished and much studied in the literature of electricity economics, it is difficult to clearly quantify a threshold or measure for unsupervised anomaly detection. The unsupervised anomaly detection task was carried out by analysing the in-sample prediction errors of LSTM-network. LSTM was suitable for the complicated multidimensional prediction task and provided comparable results. LSTM has been utilized to predict aggregated supply curves by Guo et al. (2021). Guo et al. showed that LSTM can reach validation MAPE of 3.85 in prediction of aggregated supply curves with 3.28 out-of-sample MAPE. With pre-

processing and dimension reduction the out-of-sample prediction error can reach MAPE of 2.74. In this thesis, the fold 4 provides best comparable MAPE, in which Validation MAPE is 2.41. However, results are not entirely comparable, as Guo et al. do not report how many price intervals they use in in theirs study, also the authors settled for larger number of exogenous variables. Increase of price intervals might lead to higher MAPE, but additional exogenous variables are possibly beneficial for the prediction task. Even so, LSTM can be concluded to be suitable for prediction of aggregated supply curves. In this thesis, mean absolute error over predicted volumes for every price interval of one hour was used as metric for prediction error, to provide unsupervised anomaly detection method. Qualitative survey of the prediction errors led to discoveries of actual events in the data and markets that proved unusual and in line of the research question. Unsupervised learning with LSTM algorithm can significantly benefit study of supply curves and market surveillance. Another aim in this thesis was to detect market manipulation from the hours that stood out, it proved more difficult. Price intervals chosen for this study were very sparse making it difficult to assess the true supply behaviour. Furthermore, the nature of LSTM makes statistical inference impossible, so it is not possible to say what was the fundamental reason for prediction errors. Exercise of market power can affect the supply curves in very many ways, making thorough analysis of the prediction errors difficult and tedious. The shortage of labelled hours with actual market manipulation leads to a fundamental problem in the unsupervised anomaly detection, it is difficult to train the model with non-anomalous data. It might well be true, that some manipulative behaviour in the power markets is tacit and continuous, making it hard to detect with unsupervised learning, which leans on the distribution of non-anomalous events. It should be emphasized however, that LSTM used in unsupervised learning can provide valuable information of the behaviour in the market and can be used to highlight unsuspected and unnormal events. The method in this thesis can further used to provide a dynamic threshold on market events. Further study should include more important exogenous variables such as level of hydro reservoirs, price of fuels and level of demand to the training. Also, some other price intervals could prove more practical. Further study on market manipulation should consider tacit, contextual and long-term exercise of market power. The assumption about stationary behaviour in itself includes tacit arrangements market power possessing firms potentially end up with. Prediction errors as an anomaly measure is dependent on past events in the data, and thus incorporates any long-term market power exercise in the learning process. However, the model can be used

to give information about unexpected behaviour on individual hours, which makes surveillance and further examination of bidding behaviour easier. Model would benefit from larger use of external variables, such as fuel prices, price futures and demand of electricity.

7 Conclusion

In this thesis, economical and legislative background was provided about non-genuine or abusive day-ahead market bidding behaviour. Impact of economic and physical withholding were studied on supply curves. Anomaly detection was based on assumption of stationary behaviour, which was measured as prediction error. Literature survey on studies about prediction and analysis of supply curves, as well as unsupervised anomaly detection, based the choice of a long short-term memory neural network model to be used to provide predictions of supply curves that were used as a function estimator to provide prediction errors as a measure for stationary behaviour. Model was trained in a sliding window framework to predict one hour ahead using 24 previous supply curves, time features and two weather variables. was used also to search anomalous patterns from two years of day-ahead bidding data of Finland's price area. Pre-processing included dividing supply offers to bins according to prices and standardization of the volumes in each bin to interval between zero and one. Weather data, sine and cosine curves indicating day of week time of day were used as external variables. Stacked LSTM model with two layers was chosen and grid search was used to find hyperparameters suitable for the problem. After the selection of model hyperparameters, walk forward validation was used also to collect information about the data and training process, which lead to discovery of erroneous pattern in the data. The mean absolute error of predicted and true supply curve of a hour was used as a statistic for prediction error, and as a score for anomaly. Method brought several point anomalies into attention, which were examined, and one certified case of market manipulation the algorithm detected as an anomaly was confirmed. Results were analysed quantitatively and qualitatively, few types of common errors were identified and prediction errors that could indicate capacity withholding, data error or economic withholding were labelled from the data by categorizing hours according to the intersection points between predicted and true supply curves. Although the model was capable to highlight anomalous events in supply curves promisingly, it proved difficult to make direct inference about market manipulation.

References

Abadi, M. – Agarwal, A. – Barham, P. – Brevdo, E. – Chen, Z. – Citro, C. – Corrado, Greg. S. – Davis, A. – Dean, J. – Devin, M. – Ghemawat, S. – Goodfellow, I. – Harp, A. – Irving, G. – Isard, M. – Jia, Y. – Jozefowicz, R. – Kaiser, L. – Kudlur, M. – Levenberg, J. – Mané, D. – Monga, R. – Moore, S. – Murray, D – Olah, C. – Schuster, M. – Shlens, J. – Steiner, B. – Sutskever, I. – Talwar, K. – Tucker, P. – Vanhoucke, V. – Vasudevan, V. – Viégas, F. – Vinyals, O. – Warden, P. – Wattenberg, M. – Wicke, M. – Yu, Y. – Zheng, X. (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. <http://tensor-flow.org/>, retrieved 9.9.2022

ACER (2021) *Guidance on the application of Regulation (EU) No 1227/2011 of the European Parliament and of the Council of 25 October 2011 on wholesale energy market integrity and transparency* <https://documents.acer.europa.eu/en/remit/Documents/ACER_Guidance_on_REMIT_application_6th_Edition_Final.pdf>, retrieved 22.2.2022.

ACER (2022) REMIT coordination <*https://www.acer.europa.eu/remit/coordination-on-cases/market-abuse*>, retrieved 15.2.2022.

Biggar, D – Hesamzadeh, M – Reza, M (2013) *The Economics of Electricity Markets*

Crampes, C. – Creti, A. (2006) Capacity competition in electricity markets

Chollet, Fran et al. (2015) Keras. <https://keras.io/>, retrieved 23.2.2022

Du, J. – Xu, Y. (2017) Hierarchical deep neural network for multivariate regression. *Pattern recognition*, Vol. 63, 149-157.

Energiavirasto (2021) *Päätös RAO Nordic Oyn REMIT-asetuksen markkinamanipulaatiokiellon noudattamisesta* https://energiavirasto.fi/documents/11120570/12872579/P%C3%A4%C3%A4t%C3%B6s+RAO+Nordic+Oyn+REMIT-asetuksen+markkinamanipulaatiokiellon+noudattamisesta.pdf/aa9615a6-c2f5-393e-d7e2-a9d58e28aed6/P%C3%A4%C3%A4t%C3%B6s+RAO+Nordic+Oyn+REMIT-asetuksen+markkinamanipulaatiokiellon+noudattamisesta.pdf?version=1.0&t=1592809618000, retrieved 23.2.2022

Energiavirasto (2021) *National Report. https://energiavirasto.fi/en/-/national-report-on-electricity-and-natural-gas-markets-in-2020, retrieved 22.02.2022*

Epex (2022) *Annual Report 2020.* <https://www.epexspot.com/sites/default/files/sites/annual-report-2020/markets/exchange-members-the-rise-of-the-robots/>

Fingrid (2021) 15min ISP derogation. <https://www.fingrid.fi/globalassets/dokumentit/fi/sahkomarkkinat/varttitase/final_15_min_isp_derogation_report_poyry.pdf>, retrieved 22.2.2022

Fogelberg, S. – Lasarczyk, E. (2019) Strategic Withholding through Production Failures. *The Energy journal*, Vol. 40 (1), 247.

Goodfellow, I – Bengio, Y. – Courville, A. (2016) *Deep Learning*. MIT Press, <http://www.deeplearningbook.org/>, retrieved 14.10.2021

Guo, H. – Chen, Q. – Zheng, K. – Xia, Q. – Kang, C. (2021) Forecast Aggregated Supply Curves in Power Markets Based On LSTM Model. *IEEE Transactions on Power Systems*, Vol. 36 (6). 5767-5779

Haoran,L. – Song, L. – Wang, J. – Guo, L. – Li, X. – Liang, J. (2021) Robust unsupervised anomaly detection via multi-time scale DCGANs with forgetting mechanism for industrial multivariate time series. *Neurocomputing*, Vol. 423, 444-462

Hellmer, S. – Wårell, L. (2009) On the evaluation of market power and market dominance — The Nordic electricity market.

Hundmann, K. – Colwell, I. – Constantinou, V. – Soderstrom, T. – Laporte, C. (2018) Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding

Kingma, D – Ba, J (2015) Adam: A Method for Stochastic Optimization.

Kishan, G – Chilukuri, K – HuaMing, Huang (2017) Anomaly Detection Principles and Algorithms.

Klemperer, P – Meyer, M. (1989) Supply function equilibria in oligopoly and uncertainty *Econometrica*, Vol. 57, 1243-1277.

Mansurin, E. (2008) Measuring welfare in restructured electricity markets. *The review of economics and statistics*, Vol. XC (2), 369-386.

Mehrotra, G – Mohan, C – Huang, H (2017) Anomaly Detection Principles and Algorithms.

Neuhoff, K – Barquin, J. – Boots, M.G. – Ehrenmann, A. – Hobbs, B.F. – Rijkers, F.A.M. – Vasquez, M (2005) Network-constrained Cournot models of

liberalized electricity markets: the devil is in the details, *Energy Economics*, Vol. 27 (3), 495-525

Nguyen, H.D. - Tran, K.P. - Thomassey, S - Hamad, M (2020) Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in Supply Chain Management. *International journal of information management*, Vol. 57, 102282

Nord Pool (2021) Day-ahead Trading. < https://www.nordpoolgroup.com/trading/Day-ahead-trading>, retrieved 22.2.2022

Pelegatti, M. (2013) Supply Function Prediction in Electricity Auctions.

Poletti, S (2021), Market Power in the New Zealand Electricity wholesale market 2010-2016

Rautaray, S.S. – Pemmaraju, P. – Mohanty, H. (2021) *Trends of Data Science and Applications Theory and Practices*

Tan, Y Hu, C. Zhang, K. Zheng. K Davis, E. Park, J.S. (2020) LSTM-based Anomaly Detection for Non-linear Dynamical System, IEEE access, Vol. 8, 103301-103308

Tangerås, T – Mauritzen, J. (2018) - Real-time versus day-ahead market power in hydro-based electricity market. *The Journal of industrial economics*, Vol. 66 (4), 904-941

Tran, T.N. – Le, V.D. – Dang, T.P. (2021) Grid search of multilayer perceptron based on the walkforward validation. *International Journal of Electrical and Computer Engineering.* Vol. 11 (2), 1742-1751

Varian, H (2014) *Intermediate microeconomics: a modern approach,* 9th edition.

Villerreal-Vasquez (2021) Hunting for Insider Threats Using LSTM-based Anomaly Detection. *IEEE transactions on dependable and secure computing*

Weron, R (2014) - Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting* Vol. 30 (4), 1030-1081

Willems, B – Rumiantseva, I –Weigt, H (2009) Cournot versus Supply Functions: What does the data tell us? *Energy Economics*, Vol 31 (1). 38-47

Wolak, F (2003) Measuring Unilateral Market Power in Wholesale Electricity Markets: The California Market, 1998-2000. *The American economic review,* Vol. 93 (2), 425-430

Ziel, F. – Steinert, R. (2016) Electricity price forecasting using sale and purchase curves: The X-Model. *Energy Economics*, Vol. 59, 435-454.

Chollet, Fran et al. (2015) Keras. <https://keras.io/>, retrieved 23.2.2022.

Appendices

The main heading of the appendices is not numbered. The same styles are used in the appendices as in the text chapters. The appendices are not included in the number of pages on the Abstract page.

Appendix 1 Heading

Each appendix is numbered and given a heading.

Appendix 2 Heading

You can start each appendix from a new page if you wish.