Applied Energy 319 (2022) 119182

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Empirical study of day-ahead electricity spot-price forecasting: Insights into a novel loss function for training neural networks

Ahmad Amine Loutfi [a,*], Mengtao Sun [b], Ijlal Loutfi [c,1], Per Bjarte Solibakke [d]

[a] NTNU Business School, Norwegian University of Science and Technology, NTNU Postboks 8900, Torgarden 7491, Trondheim, Norway
[b] Department of ICT and Natural Sciences, Norwegian University of Science and Technology, NTNU in Ålesund, Postboks 1517, 6025 Ålesund, Norway
[c] Department of Information Security and Communication Technology, Norwegian University of Science and Technology, NTNU in Gjøvik, Postboks 191, 2802 Gjøvik, Norway
[d] Department of International Business, Norwegian University of Science and Technology, NTNU in Ålesund, Postboks 1517, 6025 Ålesund, Norway

HIGHLIGHTS

- We proposed Theil UII-S as a novel loss function for training neural networks.
- Theil UII-S loss function provides more accurate forecasts on the average, best, and, worst case scenarios, converges faster, is twice differentiable, and has a variable gradient.
- Theil UII-S loss function is a good candidate for training neural networks for forecasting day-ahead electricity spot prices.

ARTICLE INFO

Keywords:
Machine learning
Neural networks
Loss functions
Optimization
Day-ahead forecasting

ABSTRACT

Within deregulated economies, large electricity volumes are traded in daily spot markets, which are highly volatile. To develop profitable trading strategies, all stakeholders must be empowered with robust forecasting tools. Although neural network approaches have become increasingly popular for time-series forecasting, they do not optimally capture unique features of financial datasets. A major factor hindering their performance is the choice of the backpropagation loss function. We performed a systematic and empirical study of loss functions that can optimize the forecasting of day-ahead electricity spot prices. We first outlined a set of properties that such a loss function should meet. We proposed Theil UII-S as a novel loss function, which is derived from Theil's forecast accuracy coefficient. We also implemented five neural network models and trained them on the two most used loss functions—mean squared error and mean absolute error—and our Theil UII-S. We finally tested our models on a real dataset of the electricity spot market of Norway. Our results show that Theil UII-S provides more accurate forecasts on the average, best, and, worst case scenarios, converges faster, is twice differentiable, and has a variable gradient.

1. Introduction

Before electricity markets were deregulated, spot-price predictions were laborious and detailed but straightforward. They primarily relied on estimating the market's future demand based on historical data, computing its supply by aggregating the operational costs of its available generation units, and then, comparing supply and demand values. Within such stable homogeneous markets, cost-based models—such as

Abbreviations: TSF, time series forecasting; DAM, day-ahead market; MCP, market clearing price; ReLU, rectified linear unit; SSM, state space models; MSE, mean squared error; MAE, mean absolute error; DTW, dynamic time warping; OYF, one year forward contract; OQF, one quarter forward contract; KNN, K-nearest neighbor; FFNN, feed-forward neural network; CNN, Convolutional Neural Network; RNN, recursive neural network; LSTM, long-short term memory neural network; GRU, gated recurrent unit.

* Corresponding author.
E-mail address: ahmad.a.loutfi@ntnu.no (A.A. Loutfi).

[1] The work described in this article was done while the author has been working at a postdoctoral researcher at the Norwegian University of Science and Technology. The author is now working as a product manager at Canonical, the Office Group, St Dunstans House 4th floor, 201 Borough High St London SE1 1JA, United Kingdom.

https://doi.org/10.1016/j.apenergy.2022.119182
Received 5 July 2021; Received in revised form 6 April 2022; Accepted 21 April 2022
Available online 14 May 2022

GE MAPS—were reliable tools for electricity spot-price prediction and required negligible changes to their underlying assumptions [1].

However, in today's electricity markets, such models are outdated [2,3]. Because of the increasing competition and highly volatile market conditions of today's electricity ecosystem, building prediction solutions that can accurately model the market's uncertainty and its forecast probability distribution is essential [1].

Although electricity load forecasting has advanced significantly and mature models with a mean absolute percentage error of $\leq$ 3% are available, price forecasting is still lagging [4-6]. Here, we are particularly interested in electricity spot markets in which large volumes of electricity are traded daily. Their daily volatility can reach 50%, which is 10 times that of other energy products, making their price forecasting especially challenging [7]. They are also characterized by high frequency, a variable mean, and multiple seasonalities because of a set of unique features that differentiate electricity from other forms of energy. For instance, because electricity cannot be stored, a balance between demand and supply must always be maintained [6].

Electricity spot markets can create both great opportunities and risks. To develop profitable trading strategies, all stakeholders must be empowered with appropriate price forecasting tools, which require accurate day-ahead electricity spot price forecasting models. Unfortunately, the models currently deployed and proposed are not sufficiently accurate [4,5,8].

Here, we studied day-ahead electricity spot forecasting from a technical perspective by focusing on its main building block—time series forecasting (TSF). At a high level, TSF models rely on analyzing relevant historical data to capture the underlying patterns and relationships that can help predict future behavior [9]. Given its importance in numerous real-world applications, TSF modeling has been heavily researched [10,11]. Traditionally, TSF relied on statistical state space models (SSMs), which are parametrized based on expert assumptions regarding the dataset and its domain [12]. Notable SMM examples include exponential smoothing [13], ARIMA [14], and autoregressive models [15]. Unfortunately, these models have several limitations. First, they require a priori knowledge regarding the relationship between the target and input variables, which is only feasible for a simple time series with well understood causal relationships [16,17]. Furthermore, they implicitly assume their data are linear, whereas several significant empirical studies have shown that time series within finance are primarily non-linear, using "various statistical tests" [18]. Therefore, such traditional statistical models cannot provide the most accurate forecast estimations for real-world non-linear electricity spot-price time series. To overcome their limitations, non-linear models that can learn temporal and functional relationships in an independent manner are required.

Over the last decade, data driven approaches based on advanced neural networks have been successfully employed to model various non-linear datasets within the fields of image analysis and pattern recognition [19,20]. Their main advantage is that they can inherently learn functional relationships from a dataset itself, without pre-embedding any prior assumptions or any domain expertise. This ability to model complex non-linear relationships based on a finite data sample has made neural networks a quasi-universal tool for approximating functions [16].

Inspired by their success in other fields, neural networks have been applied for time series forecasting [21]. Unfortunately, an increasing number of empirical studies have shown that, currently, neural networks still perform worse, on average, than traditional statistical models when applied to financial time series [22]. Analyzing such algorithms reveals that they fail to capture important unique features of electricity spot-price time series datasets, which, unlike the datasets of other fields, have input and output variables that are ordered in time and temporally interdependent. The electricity spot-price time series datasets comprise highly volatile data that include outliers that are prevalent by design and are not to be considered undesirable noise, as is the case in most other datasets [6].

A major factor contributing to such a performance gap is the choice of the backpropagation loss functions [23], which are borrowed from conventional problems within other fields in which neural networks were first applied (e.g., image analysis, natural language processing). The most widely employed loss functions today are the mean squared error (MSE) and mean absolute error (MAE), which both use the Jacobian matrix to compute the difference between two data points with the same abscissa [24,25]. Despite their simplicity and widespread adoption, they have several shortcomings. Although the MSE can be easily computed and converges fast even at fixed learning rates, it is sensitive to outliers because it magnifies errors by squaring them, thus reducing the overall model accuracy when the dataset has outliers. By contrast, the MAE, which is more robust to outliers, has a constant gradient, which makes it likely to miss its minima during a typical gradient descend algorithm. Furthermore, both MSE and MAE are not suitable for datasets with a wide range of values [26,27].

Given the importance of accurate time series forecasting, several studies aim to develop alternative metrics. Unfortunately, most proposed solutions have several limitations: they are either non-differentiable and thus cannot be used within the gradient descent algorithm of backpropagation; they incur a very high computational overhead; they have a constant gradient; or they do not handle outliers and noise efficiently [26-29].

In this study, we approach the problem systematically. We first outlined a set of properties that our desired loss function should satisfy. We surveyed the financial times series prediction evaluation literature and found that the Theil's forecast accuracy UII is a good potential metric around which our candidate loss function can be built [30].

The primary research question of this paper is as follows: Can Theil's forecast accuracy UII be used to build a loss function for accurately modeling day-ahead electricity spot-price time series using neural networks? This study provides a positive answer to our research question: it shows that Theil UII-S (Theil UII quadratic) provides more accurate forecasts on the average, best, and, worst case scenarios, converges faster than both MSE and MAE, is twice differentiable, its performance is independent of the chosen neural network architecture, and has a variable gradient.

To answer this research question, we performed a conceptual and an empirical analysis. First, we analyzed Theil's forecast accuracy UII against the first set of our desired properties. We then derived Theil UII-S as a candidate loss function, which we tested empirically to consolidate our theoretical findings. We designed, implemented, and deployed five neural network models with the most used architectures in time series forecasting. We then ran our models on the real-world dataset of electricity spot prices of Norway. Our results show that Theil UII-S meets the desired loss function properties and is a good candidate for training day-ahead electricity spot-price time series datasets.

The rest of this paper is organized as follows. In Section 2, we provide an overview of the relevant background concepts. In Section 3, we survey the available literature on the proposed solutions to calibrate neural networks to better forecast time series. In Section 4, we outline the properties that our desired loss function should meet and discuss Theil's forecast accuracy UII. In Section 5, we present the theoretical analysis of Theil UII and derive Theil UII-S as another candidate loss function. In Section 6, we describe our dataset and the experimental set up for our empirical study. In Section 7, we present and discuss our empirical results. Finally, we conclude the paper in Section 8 with a set of reflections upon the study's process, results, impact, and future work.

2. Background and preliminaries

2.1. Electricity markets

Ever since the early 1990 s, numerous electricity markets around the world have been undergoing progressive deregulation. This has gradually resulted in highly competitive competitive markets that are no more

under the traditional control of governments. In many ways, this shift is similar to the early development of financial markets: it is driving a growing multi-billions trading market, and creating new instruments such as spot contracts, derivatives, and futures contract [31]. However, the pricing models developed for financial instruments are not suitable for forecasting the trading prices of electricity, due to its intrinsically unique features [32]. As a matter of fact, electricity is a non-storable commodity that has no inventories. This requires new trading strategies for keeping balanced demand and supply levels. Furthermore, electricity markets are often fragmented in ways that are not economically intuitive, due to the difficulty of transferring electricity between different geographical regions, either because it is physically impossible or too expensive. Such fragmentation does not exist in financial markets: the price of Google stocks is the same in both New York and Shanghai. Moreover, electricity trading is characterized by atypical variations that also set it apart from other power commodities such as oil and gas: its supply levels are significantly impacted by highly volatile weather conditions, while its demand levels are driven by continuously changing business needs [32]. Consequently, these variations lead to unexpected price peaks, seasonalities, and variations [31].

2.2. Day-ahead spot markets and their trading mechanisms

When operating within open competitive markets, maintaining the stability of the electricity grid constantly is difficult. Thus, short term spot markets have become a crucial instrument in achieving a continuous balance between supply and demand [33]. For electricity, a successful mechanism for implementing this is day-ahead markets (DAMs), where trade occurs through an auction between interested buyers and sellers every 24 h, with many DAMs closing at 12:00 pm for all electricity to be delivered from 12:00 am and 24 h ahead. As such, DAMs augment bilateral agreements, and give the market another key opportunity to balance electricity needs that might have recently emerged [33,34].

DAM transactions are organized within trading platforms called power exchanges. All stakeholders that meet a pre-defined set of market entry requirements can connect to the power exchange and take part in its daily auctions: power generators bid with their minimum selling price and the electricity volumes they can transmit, while consumers bid with their maximum buying price and their required power volumes [33,35]. The power exchange then computes the market clearing price (MCP) as the intersection point of the supply and demand curves for each load period. The orders are then sorted and accepted according to their bid price. DAM auctions can be marginal or discriminatory. In the former, both buyers who bid above or equal to the MCP and sellers who bid below or equal to the MCP use the MPC to settle their transactions, whereas in the later, they use their biding prices [31]. Finally, to ensure the payment and delivery of all trade that takes place within the power exchange, transactions need to be cleared and settled. This role is fulfilled by the clearing house, which acts as the official proxy contractual partner between buyers and sellers [33]. Clearly, power exchanges offer numerous advantages over direct transactions: they are transparent, anonymous, mitigate risk related to payment and delivery, and give rise to one unique reference price [33].

Currently, day-ahead electricity spot markets can be found in the UK, Norway, Sweden, Spain, Finland, USA-California, the Netherlands, Germany, Denmark, Poland, and USA-Pennsylvania-New Jersey-Maryland [34].

2.3. Neural networks

Artificial neural networks—or simply neural networks—are data-driven modeling algorithms characterized by their ability to learn from a finite set of data examples [16]. They achieve this by being organized as "a massively parallel combination of simple processing units that can acquire knowledge from an environment through a learning process and store the knowledge in its connections" [36,37]. As such, neural networks are a powerful realization of the supervised learning problem, which we formally define subsequently in the context of multiple regression [38].

Let X and Y be two random variables: $X \in x \subset \mathbb{R}^d$, $Y \in y \subset \mathbb{R}$, and $Y = f(X)$ for some unknown function f. Considering a sample $\{(x_i - y_i)\}_{i=1,\ldots..n}$ drawn from the joint distribution of X and Y, supervised learning aims to learn a mapping $\widehat{f} : x \rightarrow y$ that minimizes the error, defined by a convenient loss function $L : y \times y \rightarrow \mathbb{R}$[38]. As suggested by Koushik [38], minimizing over the set of all functions from x to y is ill-posed; thus, we restrict the space of hypotheses to some set of F and define the following:

$$\widehat{f} = argminE[L(Y, f(X))], f \in F \tag{1}$$

Neural networks solve Eq. (1) by passing input x via a series of layers. Starting from the incoming input signal x, each subsequent layer is computed as follows [38]:

$$x_j = \rho W_j x_{j-1} \tag{2}$$

where W_j represents a linear operator, and ρ is a non-linear operator that makes the incoming input signal nonlinear by using an activation function such as rectified linear units (ReLU). This is an important step because the objective of neural networks is to produce a nonlinear decision boundary through combinations of weights and inputs. Each layer can thus be written as the sum of the previous layers as follows [38]:

$$x_j(u, k_i) = \rho\left(\Sigma\left(x_{j-1}(,k) * W_{j,k_j}(,k)\right)(u)\right) \tag{3}$$

In their most simple realization, neural networks aim to estimate a regression function.

$$E(Y|X = x) \tag{4}$$

by,

$$\propto + \sum_{j=1}^{h} w_j g\left(\langle \gamma^{(j)}, x > \right), \tag{5}$$

where γ is a constant, h is the bandwidth number that establishes the number of nodes, w_j is the weight of node j, g is some nonlinear function, and $\gamma^{(j)}$ denotes the weights of the variables at node j [24].

2.4. Backpropagation and optimization

The optimization problem of neural networks is highly non-convex. Typically, weight W_j is learned using a backpropagation algorithm, such as the stochastic gradient decent, which computes the gradients [38].

In this setting, backpropagation refers to the method computing the gradient and the direction that the neural network model should adopt to reduce loss, which is defined as the error difference between actual and predicted values. The model then iteratively attempts to find the minima of the loss function by changing the weights until the error converges to the lowest possible value [39].

Backpropagation is based on an expression for the partial derivative $\partial \mathrm{L}/\partial \mathrm{w}$ of the loss function L with respect to any weight w and bias b[39]. This expression describes the rate of change of the loss when weights and biases are altered. Thus, backpropagation is a simple and fast algorithm for learning and also the core mechanism that provides detailed insights into how changes in weights and biases alter the overall behavior and performance of neural networks [39].

3. Literature review

Accurate day-ahead electricity spot-price forecasts are crucial for all power portfolio managers and market stakeholders because they determine their trading strategies, as well as their production and consumption plans. This is especially true for electric utilities that often

cannot offload their costs to their retail customers; instead, when they cannot secure profitable contracts in the day-ahead markets, they take one last chance at trading within the real-time balancing markets. Unfortunately, the costs of the later are prohibitively high, and can cause substantial financial losses, even bankruptcies, during times of high volatility [31].

Although electricity prices are volatile, they are not considered random. Therefore, significant efforts are devoted toward their analysis and forecast [40]. Weron [41] classified forecasting strategies into six main categories, which outline much of the available literature in this field: 1) multi-agent models that can be based on the Nash-Cournot framework, supply function equilibrium, strategic production cost, or agent-based simulation models [31,42-45]; 2) structural models which focus on capturing the impact of the physical and economic processes that are related to electricity production and trading, such as its load and other relevant weather variables [31,46]; 3) reduced-form models which focus on risk management and the pricing of derivatives, often using jump diffusions and Markov regime switching [31,47,48]; 4) statistical models that aim to mathematically capture the potential relationships between past prices and a set of relevant past and current values of variables such as demand, production, and temperature levels [31]. Some of the most widely used statistical techniques are similar-day exponential smoothing, regression models, AR/ARX-type time series, threshold autoregressive models, and GARCH-type models [31,41,49-52]; 5) artificial intelligence models, such as neural networks and space vector machines, which learn from data without requiring a priori expert knowledge about the problem space to parametrize the model [6]; 6) hybrid approaches which combine one of more of the five aforementioned categories [31].

When applied to short-term day-ahead electricity price forecasts, statistical and artificial intelligence models are the most promising and most currently used, with the latter being a fast-growing area of research [53-55]. Among the literature on the electricity market, Anbazhagan et. al. [56] proposed a recurrent neural network based on the Elman network for forecasting day-ahead electricity spot-prices in mainland Spain. The performance results were compared to a large set of both linear and non-linear models and further tested on the electricity market of New York in 2010. Amjady et. al. [57] proposed a forecasting strategy that relies on cascaded neural networks and a novel two-stage feature selection approach. Finally, Beigaite et. al. [8] studied the Lithuanian price zone in Nord Pool by modeling its day-ahead price using seasonal naïve, exponential smoothing, and neural networks.

To overcome the shortcomings of MSE and MAE while training neural networks, several alternative loss functions have been proposed. Huber loss was proposed as a piecewise function of both MSE and MAE, in which a boundary hyper-parameter δ determines which one of the two should be used. Unfortunately, finding the right value for δ increases the complexity of training neural networks, which already have a sufficiently large number of other hyperparameters that must be finetuned [26,58]. Similarly, the log-cosh loss function computes the loss differently depending on the magnitude of the error. As a logarithm of the hyperbolic cosine of the error, loss is approximated by $\frac{1}{2}(y - f(x))^2$ when the error is large and by $|y - f(x)| - \log 2$ otherwise. However, because the gradient and Hessian of log-cosh is constant for large errors, the model's efficiency may suffer. Finally, the quantile loss function provides a forecast interval instead of a single value by assigning different penalties to negative and positive forecasts based on the value of the chosen quantile parameter γ. Smaller γ values assign larger penalties to overestimated forecasts, and vice versa. The quantile loss function is an extension of the MAE, with its value being the MAE itself when the quantile parameter is in the 50th percentile [26,59].

Within the literature, we also found several approaches that specifically focus on time series [9,10,23,24]. In [28], the authors propose replacing the backpropagation algorithm with neuro-evolutionary techniques that penalize models that do not accurately distinguish timing errors by removing them from the population pool of future models, as originally proposed by Conway et. al. [60]. While such techniques slightly improve the accuracy of short-time horizon forecasts, they significantly lag behind in terms of performance speed when compared with backpropagation-based calibration algorithms [24,28].

In [24], two intuitive correction techniques that can be applied to backpropagation optimization algorithms were proposed—error weighting and boosting. Error weighting weighs residuals during the gradient descent process and emphasizes high gradients, which are assumed to correspond to the highest timing errors. In contrast, boosting trains multiple models, and each one predicts the residuals of the previous model; their sum is used to compute the forecasted output value. Although these two techniques showed promising results when empirically used to predict flood risk in Canada, their design is strongly related to the specific characteristics of water flow datasets, and their forecast is biased toward optimizing the detection and correction of peak errors, which are correlated with the highest risk of flooding. In [9], DILATE was proposed as a novel differentiable function that optimizes the forecast of sudden changes in time series by having two distinct terms for shape and temporal changes.

Within the time series and knowledge discovery literature, we found several studies that aim to define alternative similarity measures to the Jacobian distance [23,61,62], such as the temporal distortion index [9,63,64] or the more commonly used dynamic time warping (DTW) [23]. The later was initially proposed within the speech recognition field as a general approach in which a time-wrapping function is selected first, and is then used to minimize any given loss function between two time series. Although DTW does indeed correct, to a reasonable extent, time-shift and time-wrap errors, it is computationally heavy and has no gradient, making it more suitable for pattern recognition than time series forecasting [23]. To remedy this, Frías-Paredes et. al. [63] proposed a differentiable loss function based on DTW. However, it can only be used for predicting binary time series.

4. Desired properties for the loss function

The most commonly used loss functions for training neural networks are the MSE and MAE [26]. They are both similarly used as loss functions that are minimized during the backpropagation of the optimization algorithm, namely the gradient descent. When applied to time series forecasting, MSE or MAE is first computed for each fixed-sized time interval across the series, and then, optimization is executed over the sum of all averages.

Although such loss functions have been widely successful, they are not optimal for financial time series forecasting because they do not truly capture the temporal ordered dependencies of time series datasets [9]. They also cannot provide sufficiently accurate forecasts [9]. While MAE is robust to outliers, it does so by semantically considering them as noise, which should not be de-emphasized. This is not the case for electricity spot-prices in which volatility is an inherent feature of the dataset. Furthermore, MAE has a constant gradient, as illustrated in Fig. 1, which makes it more difficult to train gradient descent algorithms, and it might miss its minima when used with gradient descent algorithms. In contrast, although MSE has a variable gradient, as shown in Fig. 2, it overemphasizes outliers, thus decreasing the overall model performance.

As discussed in Section 3, most currently proposed alternative metrics have one or several shortcomings: they are non-differentiable and thus cannot be used in backpropagation optimization algorithms, their design is biased toward optimizing the performance of specific application domains such flood control, or they are computationally heavy.

To address these shortcomings, we approach the problem systematically and outline a set of desired properties that our potential alternative loss function should meet.

Although several metrics that can accurately distinguish timing errors and properly quantify uncommon patterns have been proposed,

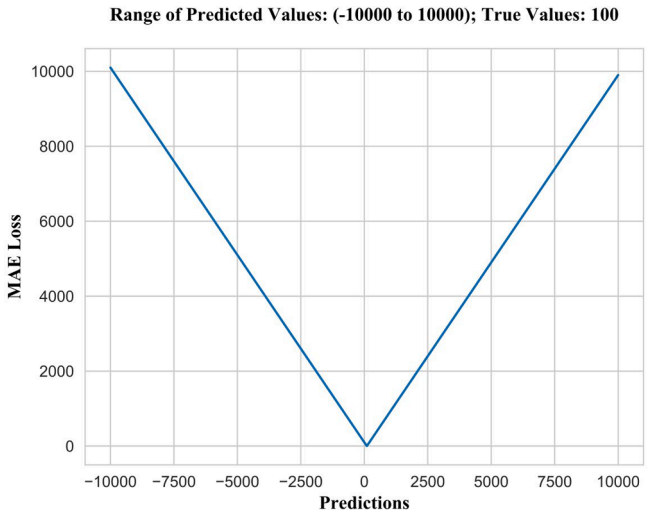

Fig. 1. Plot of MAE loss (Y-axis) vs. predictions (X-axis).

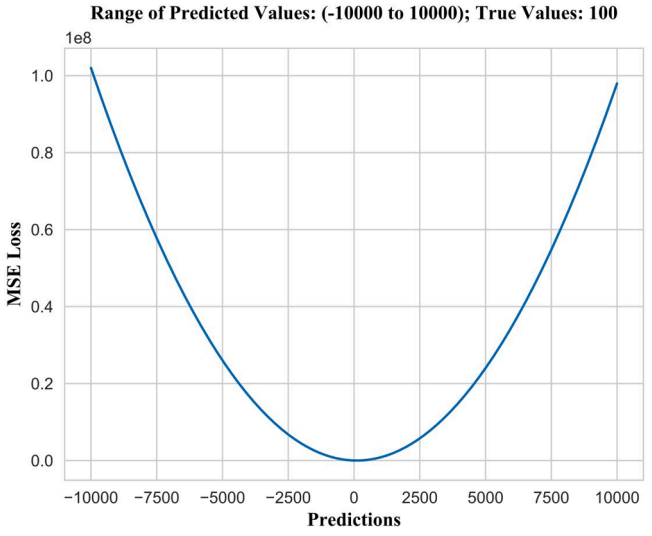

Fig. 2. Plot of MSE loss (Y-axis) vs. predictions (X-axis).

they cannot mathematically be used as a loss function, L, within a typical gradient descent algorithm. The following first property aims to address this challenge [39]:

P.1: L is differentiable with respect to the weight w and bias.b

This ensures that we can compute the partial derivative of loss L with respect to the weight w, $\partial L/\partial w$, and bias b, which is at the core of gradient descent algorithms.

Because day-ahead price forecasts are short term, the model must be trained efficiently. The next two properties address this.

P.2: L is twice differentiable.

This ensures that the loss function can leverage libraries that optimize the efficiency and portability of the gradient descent, such as the XGBoost [26,65].

P.3: L is smooth near the minima and has a variable gradient.

This will give the gradient descent algorithm a higher probability of finding the minima and ensure that it converges efficiently.

P.4: L is robust and provides accurate forecasts across different neural network architectures.

An empirical study is the most suitable approach to assess P.3 and P.4.

4.1. Proposed loss function

To find an appropriate loss function that meets the aforementioned requirements, times series prediction evaluation literature is a good place to begin. At a high level, accuracy measures can be either standalone or relative [58,60]. Although standalone measures do not require any further reference forecasts, relative ones require a benchmark forecast, relative to which the evaluation performance is measured. Because of their simplicity and general efficacy, standalone measures are widely adopted as loss functions for training neural networks, when implemented as MSE and MAE [26].

In this study, we focus on the often-overlooked class of relative measures, among which, the Theil's forecast accuracy coefficient is one of the earliest successful realizations. Formally, Theil defined two formulas that are both referred to as Theil's forecast accuracy coefficient, and which are often labeled UI and UII to distinguish them [30].

$$UI = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}f_i^2}} \tag{6}$$

$$UII = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2}{\frac{1}{n}\sum_{i=1}^{n}y_i^2}} \tag{7}$$

where y_i refers to the actual values and f_i to the corresponding predictions.

UI was suggested by Theil as a measure of forecast accuracy and is the most popular and debated coefficient among the two formulae. However, we consider UI to be unfit as a loss function for two main reasons. First, it is inconclusive and, as Theil himself later explained, "the denominator of UI depends on the forecasts and that it is therefore not true that UI is uniquely determined by the mean square prediction error" [30]. Second, its gradient is both non-symmetrical and non-smooth at the minima, as illustrated in Fig. 3.

Fortunately, an initial evaluation of the second formula of Theil's forecast accuracy coefficient UII, which was proposed as a measure of forecast quality, showed that it can potentially be a good metric for deriving our desired loss function. In fact, Theil UII is simple, clearly interpretable, and conclusive, with UII = 0 when the forecast and actual values are equal. It also unproblematically takes on a value of 1, when the standard error equals that of the naive no-change extrapolation. More interestingly, "it increases monotonically as the standard error forecasting improves over the no-change extrapolation" [30].

To solidify our intuition regarding Theil UII, we performed a

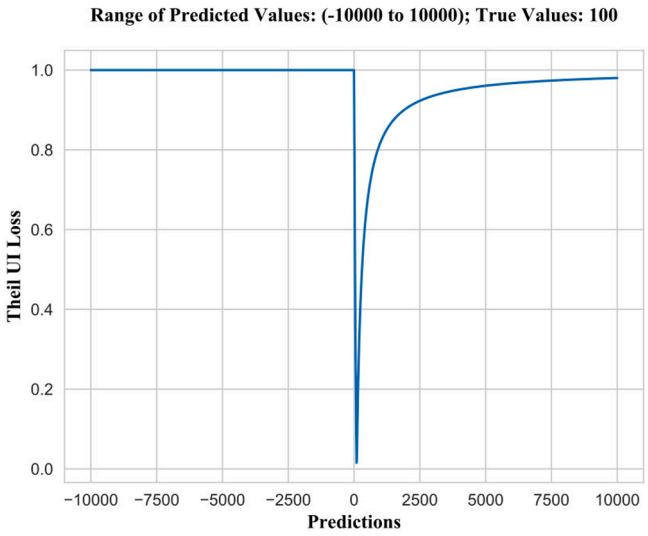

Fig. 3. Plot of Theil UI loss (Y-axis) vs. predictions (X-axis).

systemic conceptual analysis in which we evaluated it against the four properties outlined in Section 4.

5. Conceptual analysis

5.1. Theil UII

P.1: Theil UII is differentiable with respect to the weights and bias.

$$\frac{\partial Theil\ UII}{\partial w_j} = \frac{\partial}{\partial w_j}\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}$$

$$= \frac{1}{2\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}}\frac{\partial}{\partial w_j}\frac{1}{\sum_{i=1}^{n}y_i^2}\times\sum_{i=1}^{n}(y_i-f_i)^2$$

$$= \frac{1}{2\sum_{i=1}^{n}y_i^2\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}}\frac{\partial}{\partial w_j}\sum_{i=1}^{n}(y_i-f_i)^2$$

$$= \frac{1}{2\sum_{i=1}^{n}y_i^2\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}}\sum_{i=1}^{n}2(y_i-f_i)\frac{\partial}{\partial w_j}(y_i-f_i)$$

$$= \frac{1}{2\sum_{i=1}^{n}y_i^2\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}}2\sum_{i=1}^{n}(y_i-f_i)\left(\frac{\partial}{\partial w_j}y_i-\frac{\partial}{\partial w_j}f_i\right)$$

$$= \frac{-1}{\sum_{i=1}^{n}y_i^2\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}}\sum_{i=1}^{n}(y_i-f_i)\frac{\partial}{\partial w_j}f_i$$

$$= \frac{-1}{\sum_{i=1}^{n}y_i^2\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}}\sum_{i=1}^{n}(y_i-f_i)\frac{\partial}{\partial w_j}f_i \tag{8}$$

P.2: Theil UII is twice differentiable.

$$\frac{\partial^2 Theil\ UII}{\partial w_j w_k} = \frac{\partial^2}{\partial w_j w_k}\frac{-1}{\sum_{i=1}^{n}y_i^2\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}}\sum_{i=1}^{n}(y_i-f_i)\frac{\partial}{\partial w_j}f_i$$

$$= \frac{1}{2}\frac{1}{\sum_{i=1}^{n}y_i^2}\left(\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}\right)^{-\frac{3}{2}}\sum_{i=1}^{n}(y_i-f_i)\frac{\partial}{\partial w_j}f_i\sum_{i=1}^{n}(y_i-f_i)\frac{\partial}{\partial w_j}f_i$$
$$+\frac{-1}{\sum_{i=1}^{n}y_i^2\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}}\frac{\partial^2}{\partial w_j w_k}\sum_{i=1}^{n}(y_i-f_i)\frac{\partial}{\partial w_j}f_i$$

$$= \frac{1}{2}\frac{1}{\sum_{i=1}^{n}y_i^2}\left(\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}\right)^{-\frac{3}{2}}\left(\sum_{i=1}^{n}(y_i-f_i)\frac{\partial}{\partial w_j}f_i\right)^2$$
$$+\frac{-1}{\sum_{i=1}^{n}y_i^2\sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}}}\frac{\partial^2}{\partial w_j w_k}\sum_{i=1}^{n}(y_i-f_i)\frac{\partial}{\partial w_j}f_i \tag{9}$$

P.3: Theil UII is not smooth near the minima and has a variable gradient.

As illustrated in Fig. 4, Theil UII has a constant gradient that is independent of error size and is non-smooth near the minima.

5.2. Deriving Theil UII-S

To address **P.3**, we squared Theil UII and used the resulting Theil UII-S as an alternative candidate loss function:

$$Theil\ UII-S = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-f_i)^2}{\frac{1}{n}\sum_{i=1}^{n}y_i^2}$$

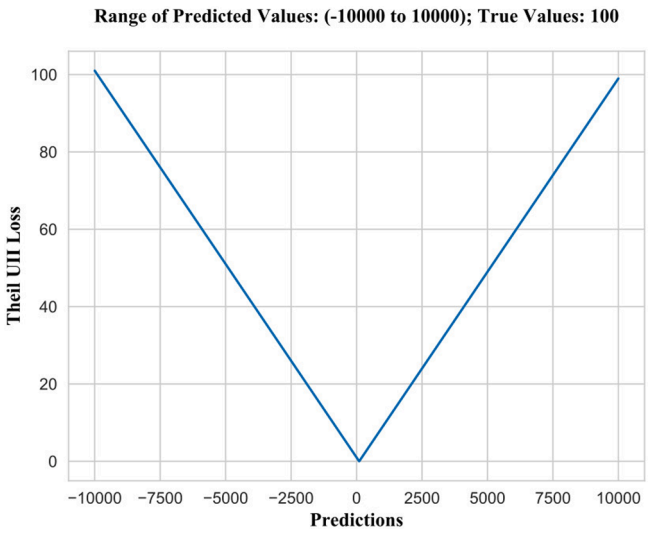

Fig. 4. Plot of Theil UII loss (Y-axis) vs. predictions (X-axis).

$$= \frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2} \tag{10}$$

Thus, Theil UII-S maintains all advantages of Theil UII, satisfies P.1 and P.2, is smooth at the minima, and has a variable gradient (P.3.). As illustrated in Fig. 5, Theil UII-S takes the smooth shape of the MSE, while its values are within the same wide range of the MAE that goes from 0 to 10000.

P.1 Theil UII-S is differentiable with respect to the weights and bias. More so, its derivative is less complicated than that of Theil UII.

$$\frac{\partial Theil\ UII-S}{\partial w_j} = \frac{\partial}{\partial w_j}\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}$$

$$= \frac{\partial}{\partial w_j}\frac{1}{\sum_{i=1}^{n}y_i^2}\times\sum_{i=1}^{n}(y_i-f_i)^2$$

$$= \frac{\sum_{i=1}^{n}\frac{\partial}{\partial w_j}(y_i-f_i)^2}{\sum_{i=1}^{n}y_i^2}$$

$$= \frac{\sum_{i=1}^{n}2(y_i-f_i)\frac{\partial}{\partial w_j}(y_i-f_i)}{\sum_{i=1}^{n}y_i^2}$$

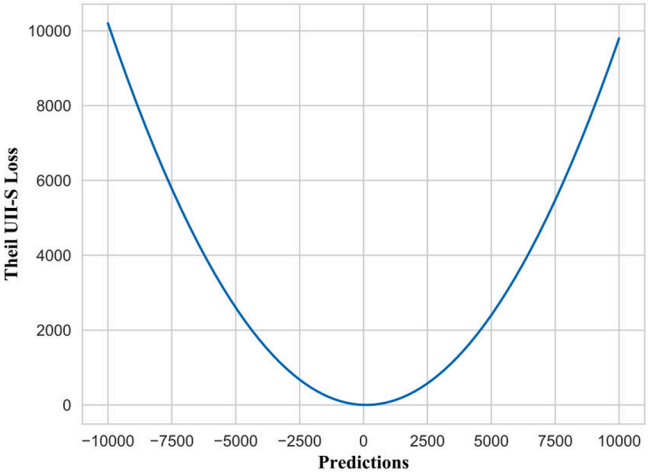

Fig. 5. Plot of Theil UII-S loss (Y-axis) vs. predictions (X-axis).

$$= \frac{2\sum_{i=1}^{n}(y_i - f_i)\left(\frac{\partial}{\partial w_j}y_i - \frac{\partial}{\partial w_j}f_i\right)}{\sum_{i=1}^{n}y_i^2}$$

$$= \frac{-2\sum_{i=1}^{n}(y_i - f_i)\frac{\partial}{\partial w_j}f_i}{\sum_{i=1}^{n}y_i^2} \quad (11)$$

Considering an activation function $g(x)$ as an inner function that takes the weighted input "net" to calculate fi, $\partial f_i/\partial w_j$ is formulated as follows:

$$\frac{\partial Theil\ UII - S}{\partial w_j} = \frac{-2\sum_{i=1}^{n}(y_i - f_i)\frac{\partial}{\partial w_j}f_i}{\sum_{i=1}^{n}y_i^2}$$

$$= \frac{-2\sum_{i=1}^{n}(y_i - f_i)\frac{\partial f_i}{\partial net_j}\frac{\partial net_j}{\partial w_j}}{\sum_{i=1}^{n}y_i^2} \quad (12)$$

P.2. Theil UII-S is twice differentiable. In addition, its second derivative is less complicated than that of Theil UII.

$$\frac{\partial^2 Theil\ UII - S}{\partial w_j w_k} = \frac{\partial^2}{\partial w_j w_k}\frac{-2\sum_{i=1}^{n}(y_i - f_i)\frac{\partial}{\partial w_j}f_i}{\sum_{i=1}^{n}y_i^2}$$

$$= \frac{2}{\sum_{i=1}^{n}y_i^2}\left(\sum_{i=1}^{n}\frac{\partial f_i}{\partial w_j}\frac{\partial f_i}{\partial w_k} - (y_i - f_i)\frac{\partial^2 f_i}{\partial w_j w_k}\right) \quad (13)$$

P.3. Theil UII is smooth near the minima and has a variable gradient. This is illustrated in Fig. 5.

P.4. To satisfy P.4. and solidify our theoretical analysis of both Theil UII and Theil UII-S, we performed an empirical study, which is detailed in Section 6.

6. Experimental setup

6.1. Data

Electricity prices have specific characteristics that differentiate them from other financial instruments and commodities [66]. "This is partly due to the inelastic short-term demand for electricity, caused by economic and business activities. Combined with the lack of efficient storage opportunities, which prevents inter-temporal smoothing of the demand, extremely large price movements (spikes) as well as various cyclical patterns of behavior occur" [66].

As illustrated in Fig. 6, spikes are infrequent and occur, for instance, because of extreme fluctuations that can be caused by extreme weather conditions, supply fluctuations due to generation outages or transmission failures, or due to the "holiday effect" [66].

We analyzed the problem of day-ahead electricity spot-price forecasting in Norway. We used historical electricity spot-price data of Norway (source NORDPOOL https://www.nordpoolgroup.com/) between January 2nd, 2013 and February 14th, 2020, as shown in Fig. 6.

Electricity spot-price (EUR/MWh) can be defined as the average market clearing price across all 24 h of the last relevant delivery day [66]. In fact, similar to other countries such as Germany, the day-ahead auction for hourly delivery of the Norwegian transmission System Operators zone occurs daily at 12.00 PM (including weekends and holidays) [66]. The market participants can submit their bids anonymously, after which the market clearing price is determined and published after 12.40 pm, with the delivery taking place during the respective hours the following day [66,67].

6.2. The dynamics of Day-ahead electricity spot-price forecasting

In this section, we aim to analyze the underlying mechanism that influences day-ahead spot-prices, and to motivate our choice of the model's input variables.

The supply and demand levels drive market prices. For electricity day-ahead markets where the grid must always be balanced, its demand must always equal its supply [66]. However, modeling real supply and demand levels alone is not enough. This is because electricity spot-prices are mainly driven by the expectations of the market about supply and demand, and not by their real volumes [68]. Therefore, our price forecasting model needs to accurately capture the predicted levels of supply and demand.

On the demand side, Norway has registered a yearly increase in power consumption between 1 and 1.5 TWh [68]. While such capacities can meet the demands of at least 50 000 homes per year, the Norwegian market electricity consumption growth rate has been 5 times larger than that of electricity supply over the last ten years [68]. Therefore, the Norwegian electricity market can expect prices that will continue to rise. Furthermore, electricity demand levels are greatly correlated to the strength of the business activity in other markets, such as commodities [68]. Instead of reinventing the wheel and building forecast models for such markets, we instead use the value of the forward electricity contracts that they issue, as they are based on their future forecasted activity levels, and thus, their electricity needs (see Table 1).

On the supply side, the expected production volume of the Nordic electricity market is directly correlated to its supply capacity, and thus, its spot-prices. For instance, since developing new Nordic power plants

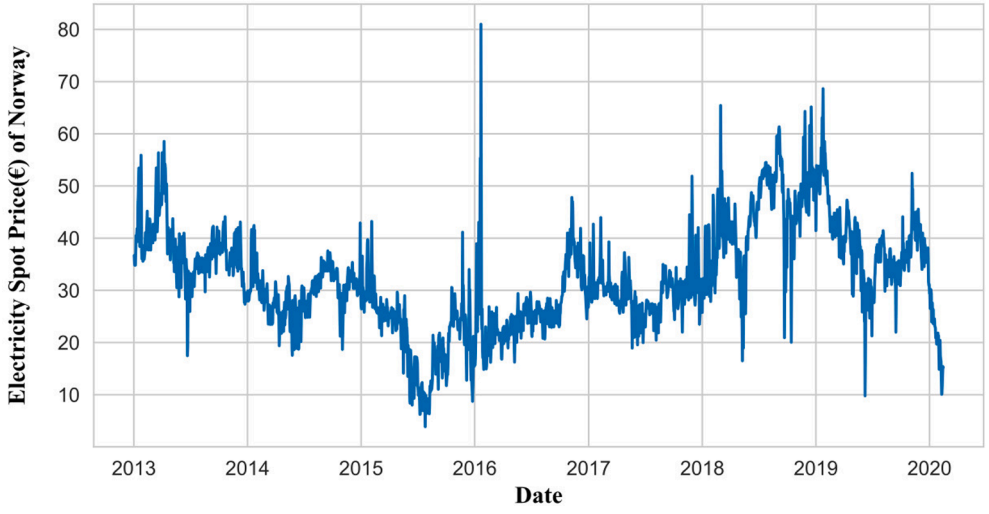

Fig. 6. Line chart of electricity spot price of Norway in 2013–2020.

Table 1
List of the input variables used in the analysis and their sources.

Input variables	Source
Electricity consumption prognosis, MWh	Leading power market in Europe NORDPOOL https://www.nordpoolgroup.com/
Electricity production prognosis, MWh	NORDPOOL
Wind prognosis, MWh	NORDPOOL
One quarter forward contract price, EUR	NORDPOOL
One year forward contract price, EUR	NORDPOOL
Brent oil price prognosis, EUR	Bloomberg, Ticker: COA Comdty
Coal price, EUR	European Energy Exchange https://www.eex.com/en/

is forecast to fall between 0.25 and 0.030 NOK per kWh, we can realistically expect future electricity prices not to exceed this price range, if we assume that extending production volumes will also remain realistically feasible whenever new demand levels arise [68]. Brent oil is generally a large contributor to primary energy production. While its use in Norway has been recently decreasing due to the current environmental legislation, oil prices still significantly impact the transportation cost of other portable renewable energies [66,67]. As such, oil impacts the electricity spot price in Norway owing to its contribution to both its production and transportation. Coal is another important driver of electricity day-ahead spot markets. As European plants need to buy raw coal from other third parties, they can sometimes halt their production when rising coal prices make it unprofitable to burn. In turn, the decreasing coal production levels decrease the overall electricity supply and increase its spot-price [68]. Last but not the least, renewable energies play an increasingly important role in today's electricity production, largely because of their near-zero carbon emissions [66]. This is especially true for wind power where production costs have been dramatically decreasing for several years, and where the cost of setting up a new wind farm has already become cheaper than that of coal plants or gas in many countries [69]. Being one of the fastest growing global energies, wind needs to be part of any robust day-ahead electricity spot-prices forecasting model.

6.3. Data preparation

6.3.1. Initial exploration

An initial exploration of the dataset was conducted through a statistical description of the input variables, as detailed in Table 2. For instance, some input variables contained missing values, which are mainly due to the markets being closed during weekends and public holidays. In Section 6.3.3, we explain how we deal with the missing values. Furthermore, the descriptive statistics show that values of input variables have different scales. While neural networks are known for their ability to process raw data, different studies have shown that normalizing data enhances convergence and generalization in most tasks [70]. Therefore, all our data were normalized in the same range, with a mean of 0 and a standard deviation of 1.

6.3.2. Missing values

Missing data values are a major concern in machine learning because knowledge is primarily extracted from data and largely depends on its quality and completeness [71]. Therefore, identifying all missing values within our dataset and handling them before feeding data into the neural network are important tasks. In this study, we deployed five different techniques to deal with missing values and compared their results before finally selecting the most optimal one for our dataset:

Ignore: we simply ignore the missing values by deleting them. Despite its simplicity, this technique can make data more biased, which will then negatively impact the model's performance [71-73].
Mean: the missing values are replaced with the mean of the weekly electricity prices.
Cubic spline interpolation: this mathematical technique constructs new data points within the boundaries of a set of other known points. These new data points are output by the interpolation/spline function, which consists of multiple cubic piecewise polynomials [74].

Given a set of $n+1$ data points $(x_i - y_i)$ where no two x_i are the same and $a = x_0 < x_1 \ldots\ldots x_n = b$, the spline $S(x)$ is a function satisfying the following conditions:

- o $S(x) \in C^2[a,b]$:
- o On each subinterval $[x_{i-1}, x_i], S(x)$ is a polynomial of degree 3, where $i = 1, \ldots, n$.
- o $S(x_i) = y_i, forall\ i = 0, 1, \ldots, n$.
- o Nearest value: the missing values are replaced with those of the nearest day.
- o K-nearest neighbor (KNN): KNN is a part of the family of hot deck imputation methods; it replaces missing values with values extracted from donors that are similar to the recipient as per a given similarity measure [71]. In this study, we set k as 7 and replaced the missing values with the average of their seven nearest measured data points. The choice of K = 7 is based on the weekly trend that can characterize electricity spot-price datasets.

To select the seven nearest neighbors, we used the Euclidean norm as a similarity measure. It assumes the following form:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{14}$$

where p and q refer to two points in the Euclidean n-space, and p_i and q_i are Euclidean vectors starting from the origin of the space (initial point).

To assess the potential impact of the above five data imputation approaches on the forecasting performance, we tested them on a feed-forward neural network (FFNN). We did not change the model's configuration to isolate the impact of each missing value on the forecasting accuracy. As shows in Table 3, KNN performs best on our dataset. Therefore, we used it to fill the values of our missing data points in the final experimental design.

Table 2
Descriptive statistics of the input variables and target variables between January 2nd, 2013 and February 14th, 2020.

Input variables	Electricity spot price	Electricity consumption prognosis	Electricity production prognosis	Wind prognosis	One quarter forward contract price	One year forward contract price	Brent oil price prognosis	Coal price prognosis
Count	2600	2600	2600	2600	1784	1784	1830	1834
Mean	32.4	1,126,639	1,096,378	78833.6	34.8	32.2	70.1	67.8
Median	31.3	108,578	1,068,124	68,729	34.7	33.7	63.6	65.1
Std Dev	9.93	193,890	193,455	50,941	10.32	7.1	22.26	14.1
Minimum	3.9	679,182	668,090	2409	13.4	16.3	27.9	46.9
Maximum	81	1,663,751	1,614,981	292,026	58.7	47.5	115.1	102.6

Table 3
Impact of the missing value technique on the forecasting accuracy using a FFNN.

Missing value technique	MSE	MAE
Ignore	1.35	3.05
Mean	1.19	1.84
Nearest value	1.30	2.23
KNN (K-nearest neighbor)	**0.84**	**1.00**
Cubic spline	1.48	3.03

6.3.3. Study of correlations

To solidify our choice of input variables for the dataset, the correlation between its different features (target and explanatory) was explored by computing the Pearson, Spearman, and Kendall correlation matrices (see Figs. 7-9).

Overall, the three correlation matrices show that the input variables and electricity spot-price are positively correlated. The Pearson, Spearman, and Kendall matrices showed slightly different correlations between the input variables and electricity spot-price. We decided to keep them as part of our dataset because we attribute the observed differences to the sensitivity acceptance of each method.

7. Experimental results

7.1. Model development

Our dataset includes the consecutive recordings of 2600 days from January 2nd, 2013 to February 14th, 2020. After the preprocessing phase, the dataset was divided into a training set (the first 1600 days), validation set (the next 400 days), and test set (the last 600 days). To ensure that our results were robust and not specific to any one neural network architecture, we developed five models to test our loss function: FFNN, convolutional neural network (CNN), recursive neural network (RNN), long-short term memory neural network (LSTM), and gated recurrent unit (GRU) neural network. The code and data are available in the Git hub depository: https://github.com/ahmadamineloutfi.

Our design follows the principles of simplicity, where we keep each model at 1 hidden layer with 64 corresponding neurons. As the activation function, we primarily used ReLU. We also used the RMSprop as the optimization algorithm for the models' stochastic gradient descent.

7.2. Performance of Theil UII and Theil UII-S

We trained all neural network models on four loss functions—MSE, MAE, Theil UII, and Theil UII-S—which gave us a total of 20 models. We measured the performance of each model based on both MSE and MAE. To understand the real impact of the proposed loss functions on forecast accuracy, the hyperparameters were the same for all models—learning rate, activation function, optimizer, batch size, and lookback interval.

Furthermore, given the strong relationship between the loss function and the model's lookback and performance, we allowed each model to run on a range of lookback values from 1 to 12. This is to allow a fair comparison of the results, as each neural network and loss function is expected to perform best on a different lookback.

As such, we present our results for the test set as follows:

Average forecast analysis: For each model, we computed the mean forecast error across all lookbacks as well as its standard deviation. The results presented in Table 4 clearly show that Theil UII-S outperforms both MSE and MAE as a loss function in 10/10 of the test cases because it gives the lowest average error across all 12 lookbacks when evaluated with both MSE and MAE. In other words, if we choose a model and a lookback value randomly, then training it on Theil UIIS will give more accurate forecasts on average. Furthermore, the small standard deviation on the Theil UII-S measure of 0.37 with the MSE and 0.14 with the MAE shows that it fulfills **P.4**:

P.4: L is robust and provides accurate forecasts across different neural network architectures.

We can see that Theil UII provides less accurate forecasts than Theil UII-S. Nonetheless, Theil UII still outperforms MSE in 80% test models and MAE in 60% test models.

Optimal forecast analysis: For each model and each loss function, we only kept the lookback results that correspond to the best forecast accuracy (see Fig. 10). The results presented in Table 5 show that Theil UII and Theil UII-S still outperform both MSE and MAE in 9/10 test models, out of which Theil UII-S provides the best forecasts 60% of the

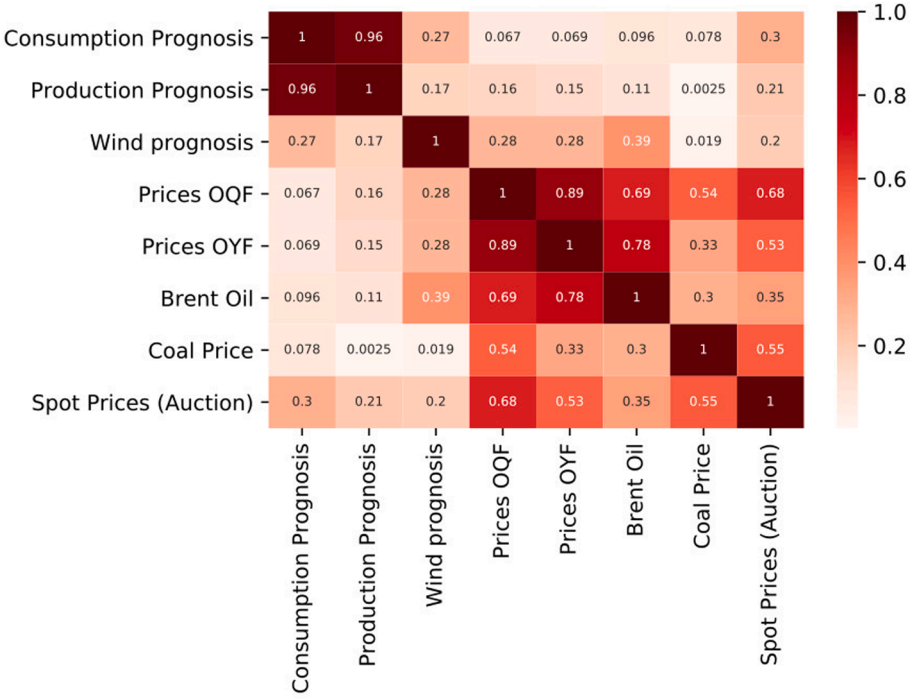

Fig. 7. Pearson correlation matrix.

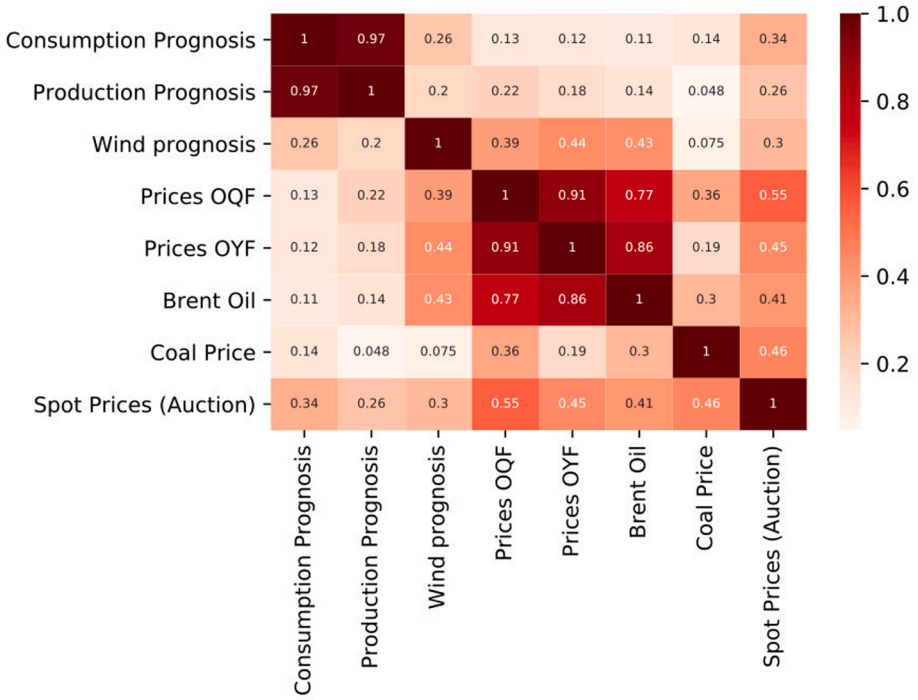

Fig. 8. Spearman correlation matrix.

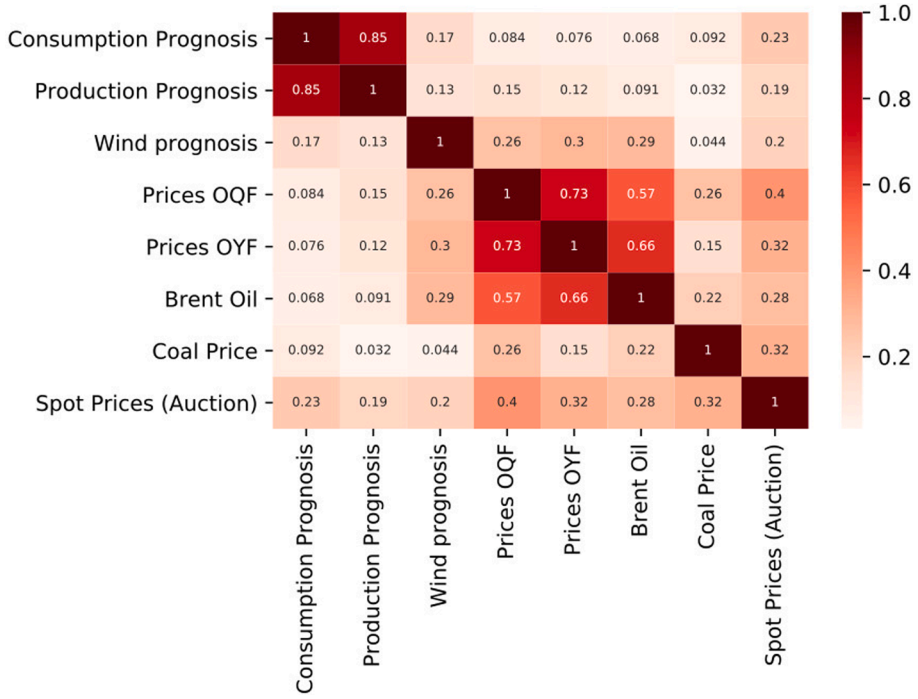

Fig. 9. Kendall correlation matrix.

time.

Worst Forecast analysis: For each model and each loss function, we only kept lookback results that correspond to the worst forecast accuracy. The results presented in Table 6 show that Theil UII-S still outperforms both MSE and MAE in 10/10 of the test cases, with upper bound errors of 2.27 and 1.31 as measured by the MSE and MAE, respectively.

7.3. Convergence

Figs. 11 and 12 show a comparison of the number of epochs required for each loss to reach its minimum error. They clearly show that neural networks trained on Theil UII and Theil UII-S converge significantly faster than those trained on MSE and MAE. This is true for most of the implemented test models on the 12 different lookbacks.

Table 4
Overview of the average forecast analysis of the loss functions used in training five different neural network architectures on 12 lookbacks.

	FFNN		CNN		LSTM		GRU		RNN	
Evaluation metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
MSE	1.48 ± 0.82	0.89 ± 0.33	1.86 ± 0.86	1.28 ± 0.70	1.19 ± 0.80	0.80 ± 0.26	0,85 ± 0.35	0.74 ± 0.16	1.03 ± 0.34	0.85 ± 0.21
MAE	1.53 ± 0.86	0.97 ± 0.25	2.11 ± 0.90	1.10 ± 0.20	1.06 ± 0.98	0.72 ± 0.30	1.10 ± 0.37	0.89 ± 0.21	1.56 ± 1.02	0.92 ± 0.23
Theil UII	*1.43 ± 0.69*	1.00 ± 0.39	*1.53 ± 0.53*	*0.99 ± 0.21*	*1.10 ± 0.63*	0.80 ± 0.28	*0.81 ± 0.31*	*0.72 ± 0.10*	1.26 ± 0.57	*0.91 ± 0.23*
Theil UII-S	**1.06 ± 0.34**	**0.82. ± 0.11**	**1.23 ± 0.51**	**0.88 ± 0.20**	**0.76 ± 0.30**	**0.64 ± 0.13**	**0.63 ± 0.35**	**0.61 ± 0.09**	**0.72 ± 0.33**	**0.64 ± 0.17**

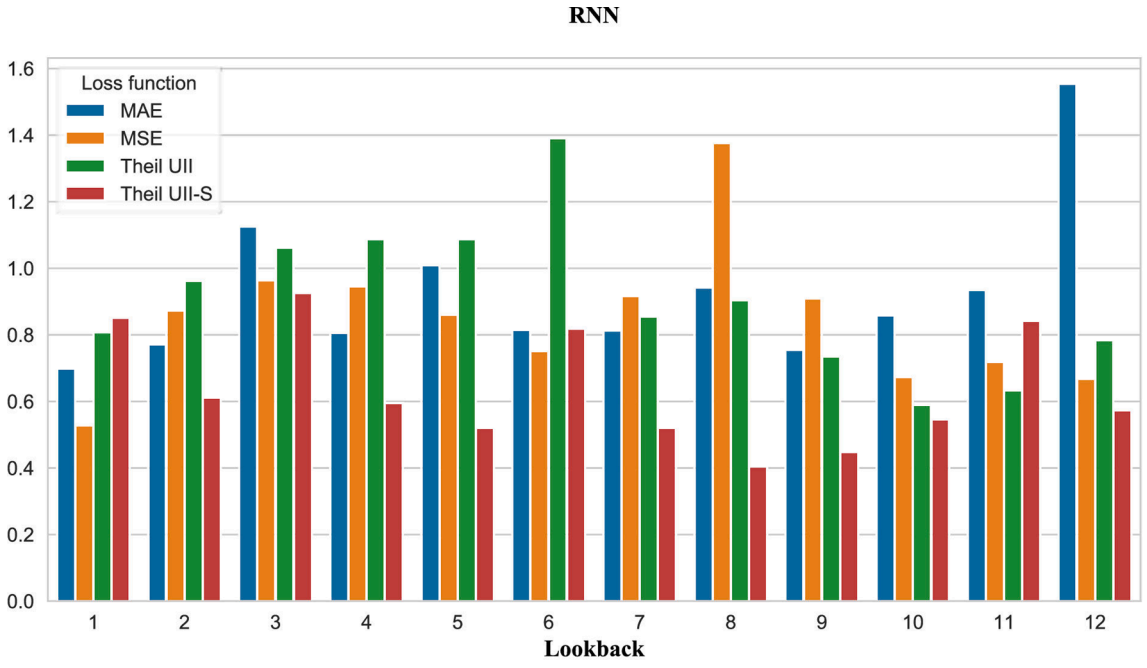

Fig. 10. RNN results using four different loss functions on 12 lookbacks and MAE as an evaluation metric.

Table 5
Overview of the optimal forecast analysis of the loss functions used in training five different neural network architectures on 12 lookbacks.

Evaluation metric	MSE			MAE		
	Loss function	Lookback	Value	Loss function	Lookback	Value
FFNN	Theil UII	4	0,29	MSE	8	0,32
CNN	**Theil UII-S**	**10**	**0,41**	**Theil UII-S**	**11**	**0,52**
LSTM	Theil UII	1	0,33	Theil UII	1	0,44
GRU	**Theil UII-S**	**8**	**0,35**	**Theil UII-S**	**9**	**0,50**
RNN	**Theil UII-S**	**8**	**0.32**	**Theil UII-S**	**8**	**0,40**

8. Summary of results

Our empirical results show that models trained on Theil UII-S provide more accurate forecasts for day-ahead electricity prices than those trained on MSE, MAE, and Theil UII. This is true for the average, best-case, and worst-case scenarios. The performance of Theil UII-S is also independent of the specific neural network architecture and converges the fastest.

Although Theil UII also performed reasonably well, it was outperformed by Theil UII-S in most test cases. We would argue that this is because Theil UII-S gives more weight to "outliers" than Theil UII because Theil UII-S squares their errors. While this would have decreased the model's performance in other datasets, it instead improves the day-ahead electricity spot-price forecasts because outliers are an inherent feature of the data; they represent its high volatility and unusual peaks that must not be minimized as noise.

Considering the empirical results as well as the conceptual analysis, we conclude that Theil UII-S meets all the desired properties outlined in Section 4, and that it is indeed a good candidate loss function for training day-ahead electricity spot-prices.

9. Conclusions and future work

In this study, we aimed to solve one of the problems of day-ahead electricity spot-price forecasting systematically- the choice of the backpropagation loss function. First, we provided a set of desirable properties that such a loss function should meet. We then proposed Theil UII-S as a novel loss function based on Theil's forecast accuracy coefficient UII. We analyzed Theil UII-S (Theil UII quadratic) conceptually and found that it meets the first set of our outlined properties. To solidify our initial findings, we conducted an empirical study in which we trained five neural networks on MSE, MAE, Theil UII, and Theil UII-S. We finally tested our models on the real dataset of electricity spot price in Norway. Our results show that Theil UII-S provides more accurate forecasts on the average, best, and, worst case scenarios. We also found that it converges faster than both MSE and MAE, is twice

Table 6
Overview of the worst forecast analysis of the loss functions used in training five different neural network architectures on 12 lookbacks.

Evaluation Metrics	MSE				MAE			
Loss Function	**MSE**	**MAE**	**Theil UII**	**Theil UII-S**	**MSE**	**MAE**	**Theil UII**	**Theil UII-S**
FFNN	3.06	3.44	2.71	**1.67**	1.39	1.48	1.92	**0.99**
CNN	3.46	3.77	3.05	**2.27**	3.41	1.46	1.48	**1.31**
LSTM	3.18	3.94	2.14	**1.50**	1.34	1.59	1.28	**0.90**
GRU	1.63	1.72	1.46	**1.67**	1.14	1.48	0.86	**0.76**
RNN	1.47	4.56	2.39	**1.21**	1.38	1.55	1.39	**0.93**

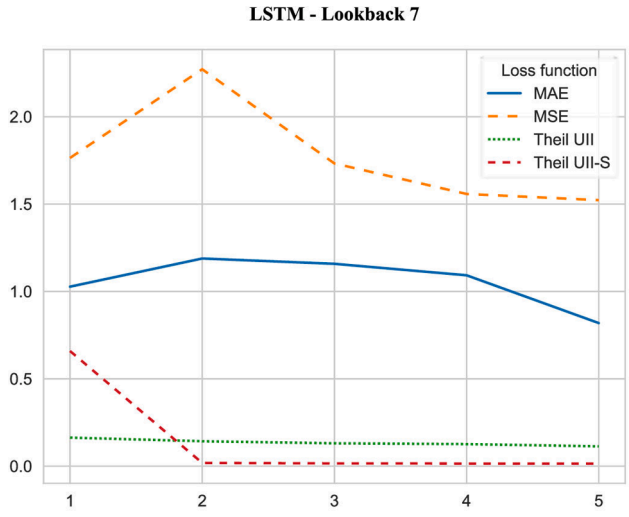

Fig. 11. Loss function convergence in LSTM.

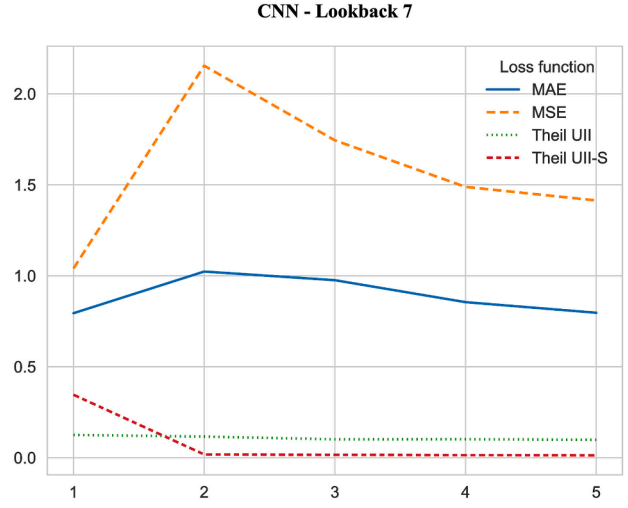

Fig. 12. Loss function convergence in CNN.

differentiable, its performance is independent of the chosen neural network architecture, and it has a variable gradient. Therefore, we conclude that Theil UII-S is indeed a good candidate for training neural networks in forecasting day-ahead electricity spot prices.

Despite its many advantages, this study has some limitations, such as using only one real-world dataset for verification. As future work, we plan to empirically test Theil UII-S on more financial time series datasets to draw stronger conclusions about its generalizability. We also aim to study whether Theil UII-S can accurately distinguish temporal and amplitude errors.

CRediT authorship contribution statement

Ahmad Amine Loutfi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Mengtao Sun:** Methodology, Software, Investigation, Resources, Data curation, Writing – review & editing, Visualization. **Ijlal Loutfi:** Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Per Bjarte Solibakke:** Resources, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Angelus A. Electricity Price Forecasting in Deregulated Markets. Electr J 2001;31 (1). https://doi.org/10.1016/j.ijepes.2008.09.003.
[2] Bye T, Hope E. Deregulation of electricity markets: the Norwegian experience. Econ Polit Wkly 2005.
[3] Arango S, Dyner I, Larsen ER. Lessons from deregulation: Understanding electricity markets in South America. Util. Policy 2006;14(3):196–207. https://doi.org/10.1016/j.jup.2006.02.001.
[4] Abdel-Aal RE. Modeling and forecasting electric daily peak loads using abductive networks. Int J Electr Power Energy Syst 2006;28(2):133–41. https://doi.org/10.1016/j.ijepes.2005.11.006.
[5] Mandal P, Senjyu T, Urasaki N, Funabashi T. A neural network based several-hour-ahead electric load forecasting using similar days approach. Int J Electr Power Energy Syst 2006;28(6):367–73. https://doi.org/10.1016/j.ijepes.2005.12.007.
[6] Aggarwal SK, Saini LM, Kumar A. Electricity price forecasting in deregulated markets: A review and evaluation. Int J Electr Power Energy Syst 2009;31(1): 13–22. https://doi.org/10.1016/j.ijepes.2008.09.003.
[7] Weron R, Misiorek A. Forecasting spot electricity prices with time series models. In: In: Proceedings of the European electricity market EEM-05 conference; 2005. https://doi.org/10.1016/j.ijforecast.2008.08.004.
[8] Beigaite R, Krilavičius T, Man KL. Electricity Price Forecasting for Nord Pool Data. In: In: 2018 International Conference on Platform Technology and Service (PlatCon); 2018. https://doi.org/10.1109/PlatCon.2018.8472762.
[9] Guen V, Thome N. Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models. arXiv:1909.09020v4 [stat.ML] 2019.
[10] Fan C, Zhang Y, Pan Y, Li X, Zhang C, Yuan R, et al. Multi-horizon time series forecasting with temporal attention learning. In: In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019. https://doi.org/10.1145/3292500.3330662.
[11] Cinar YG, Mirisaee H, Goswami P, Gaussier E, Aït-Bachir A, Strijov V. Position-based content attention for time series forecasting with sequence-to-sequence rnns. In: International conference on neural information processing 2017.
[12] Lim B, Zohren S. Time Series Forecasting With Deep Learning: A Survey. Phil Trans R Soc 2021;379(2194):20200209. https://doi.org/10.1098/rsta.2020.0209.
[13] De Livera AM, Hyndman RJ, Snyder RD. Forecasting time series with complex seasonal patterns using exponential smoothing. J Am Stat Assoc 2011;106(496): 1513–27.
[14] Chen P, Pedersen T, Bak-Jensen B, Chen Z. ARIMA-based time series model of stochastic wind power generation. IEEE Trans Power Syst 2010;25(2):667–76. https://doi.org/10.1109/TPWRS.2009.2033277.
[15] Lewis R, Reinsel GC. Prediction of multivariate time series by autoregressive model fitting. J Multivar Anal 1985;16(3):393–411. https://doi.org/10.1016/0047-259X(85)90027-2.
[16] Sánchez-Sánchez PA, García-González JR, Coronell LHP. Encountered Problems of Time Series with Neural Networks: Models and Architectures. Recent Trends Artif Neural Netw - Train Predict 2019.

[17] Qi M, Zhang GP. An investigation of model selection criteria for neural network time series forecasting. Eur J Oper Res 2001;132(3):666–80. https://doi.org/10.1016/S0377-2217(00)00171-5.
[18] Franses PH, van Dijk D. Nonlinear time series models in empirical finance. Cambridge, UK; New York: Cambridge University Press; 2000.
[19] Egmont-Petersen M, de Ridder D, Handels H. Image processing with neural networks—a review. Pattern Recognit 2002;35(10):2279–301. https://doi.org/10.1016/S0031-3203(01)00178-9.
[20] Goldberg Y. Neural network methods for natural language processing. Synth Lect Hum Lang Technol 2017;10(1):1–309. https://doi.org/10.2200/S00762ED1V01Y201703HLT037.
[21] Tealab A. Time series forecasting using artificial neural networks methodologies: A systematic review. Future Comput Inform J 2018;3(2):334–40. https://doi.org/10.1016/j.fcij.2018.10.003.
[22] Khan MY. Advances in applied nonlinear time series modeling. Ludwig Maximilian University of Munich; 2015.
[23] Rivest F, Kohar R. A New Timing Error Cost Function for Binary Time Series Prediction. IEEE Trans Neural Netw Learn Syst 2020;31(1):174–85. https://doi.org/10.1109/TNNLS.2019.2900046.
[24] Snieder E. Artificial neural network-based flood forecasting: Input variable selection and peak flow prediction accuracy. York University 2019.
[25] Seibert SP, Ehret U, Zehe E. Disentangling timing and amplitude errors in streamflow simulations. Hydrol Earth. Syst Sci 2016;20(9):3745–63. https://doi.org/10.5194/hess-20-3745-2016.
[26] Wang Qi, Ma Y, Zhao K, Tian Y. A Comprehensive Survey of Loss Functions in Machine Learning. Ann Data Sci 2022;9(2):187–212. https://doi.org/10.1007/s40745-020-00253-5.
[27] Grover P. 5 Regression Loss Functions All Machine Learners Should Know [accessed 6 April 2021] Medium 2020. https://heartbeat.fritz.ai/5-regression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0.
[28] Abrahart RJ, Heppenstall AJ, See LM. Timing error correction procedure applied to neural network rainfall—runoff modelling. Hydrol Sci J 2007;52(3):414–31. https://doi.org/10.1623/hysj.52.3.414.
[29] Itakura F. Minimum prediction residual principle applied to speech recognition. IEEE Trans Acoust Speech. Signal Process 1975;23(1):67–72. https://doi.org/10.1109/TASSP.1975.1162641.
[30] Bliemel F. Theil's Forecast Accuracy Coefficient: A Clarification. J Marketing 1973;10(4):444. https://doi.org/10.2307/3149394.
[31] Weron R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. Int J Forecast 2014;30(4):1030–81. https://doi.org/10.1016/j.ijforecast.2014.08.008.
[32] Geman H, Dauphine U.P.I. Towards a European Market of Electricity: Spot and Derivatives Trading. University Paris IX Dauphine and ESSEC 2002.
[33] EpexSpot. Basics of the Power Market, https://www.epexspot.com/en/basicspowermarket ; [accessed 01st February 2022].
[34] Acaroğlu H, García Márquez FP. Comprehensive Review on Electricity Market Price and Load Forecasting Based on Wind Energy. Energies 2021;14(22):7473. https://doi.org/10.3390/en14227473.
[35] Von der Fehr NH, Harbord D. Competition in Electricity Spot Markets: Economic Theory and International Experience. Oslo: Univ; 1998.
[36] Haykin S. Neural Networks: A comprehensive foundation. 2nd edition. Pearson Education; 2004.
[37] Guresen E, Kayakutlu G. Definition of artificial neural networks with comparison to other networks. Procedia Comput Sci 2011;3:426–33. https://doi.org/10.1016/j.procs.2010.12.071.
[38] Koushik J. Understanding Convolutional Neural Networks. arXiv:1605.09081v1 [stat.OT] 2016.
[39] Nielsen MA. Neural networks and deep learning. Determination press; 2015.
[40] Monroy JJR, Kita H, Tanaka E, Hasegawa J. Price forecasting in the day-ahead electricity market. In 39th International Universities Power Engineering Conference. 2004.
[41] Weron R. Modeling and forecasting electricity loads and prices: A statistical approach. John Wiley & Sons; 2007.
[42] Day CJ, Hobbs BF, Jong-Shi Pang. Oligopolistic competition in power networks: a conjectured supply function approach. IEEE Trans Power Syst 2002;17(3):597–607. https://doi.org/10.1109/TPWRS.2002.800900.
[43] Bolle F. Competition with supply and demand functions. Energy Econ 2001;23(3):253–77. https://doi.org/10.1016/S0140-9883(00)00061-X.
[44] Batlle C, Barquín J. A strategic production costing model for electricity market price analysis. IEEE Trans Power Syst 2005;20(1):67–74. https://doi.org/10.1109/TPWRS.2004.831266.
[45] Guerci E, Rastegar MA, Cincotti S. Agent-based modeling and simulation of competitive wholesale electricity markets. In: Handbook of power systems II. Springer; 2010. https://doi.org/10.1007/978-3-642-12686-4_9.
[46] Gonzalez V, Contreras J, Bunn DW. Forecasting power prices using a hybrid fundamental-econometric model. IEEE Trans Power Syst 2012;27(1):363–72. https://doi.org/10.1109/TPWRS.2011.2167689.
[47] Albanese C, Lo H, Tompaidis S. A numerical algorithm for pricing electricity derivatives for jump-diffusion processes based on continuous time lattices. Eur J Oper Res 2012;222(2). https://doi.org/10.2139/ssrn.1018493.
[48] Christensen T, Hurn S, Lindsay K. It never rains but it pours: modeling the persistence of spikes in electricity prices. Energy J 2009;30(1).
[49] Tong h. Non-linear time series: a dynamical system approach. Oxford university press, 1990.
[50] Kim C, Yu IK, Song YH. Prediction of system marginal price of electricity using wavelet transform analysis. Energy Convers Manag 2002;vol. 43, no. 14.
[51] Crespo Cuaresma J, Hlouskova J, Kossmeier S, Obersteiner M. Forecasting electricity spot-prices using linear univariate time-series models. Appl Energy 2004;77(1):87–106. https://doi.org/10.1016/S0306-2619(03)00096-5.
[52] Knittel CR, Roberts MR. An empirical examination of restructured electricity prices. Energy Econ 2005;27:5.
[53] Zaroni D, Piazzi A, Tettamanti T, Sleisz A. Investigation of Day-ahead Price Forecasting Models in the Finnish Electricity Market. In Proceedings of the 12th International Conference on Agents and Artificial Intelligence, Valletta, Malta, 2020.
[54] Singhal D, Swarup KS. Electricity price forecasting using artificial neural networks. Int J Electr Power Energy Syst 2011;33(3):550–5. https://doi.org/10.1016/j.ijepes.2010.12.009.
[55] Szkuta BR, Sanabria LA, Dillon TS. Electricity price short-term forecasting using artificial neural networks. IEEE Trans Power Syst 1999;14(3):851–7. https://doi.org/10.1109/59.780895.
[56] Anbazhagan S, Kumarappan N. Day-Ahead Deregulated Electricity Market Price Forecasting Using Recurrent Neural Network. IEEE Syst J 2013;7(4):866–72. https://doi.org/10.1109/JSYST.2012.2225733.
[57] Amjady N, Keynia F. Day-ahead price forecasting of electricity markets by a new feature selection algorithm and cascaded neural network technique. Energy Convers Manag 2009;50(12):2976–82. https://doi.org/10.1016/j.enconman.2009.07.016.
[58] Huber PJ. Robust estimation of a location parameter. Annals of Mathematical Statistics 1992;35. https://doi.org/10.1007/978-1-4612-4380-9_35.
[59] Koenker R, Hallock KF. Quantile regression. J Econ Perspect 2001;15(4):143–56. https://doi.org/10.1257/jep.15.4.143.
[60] Conway AJ, Macpherson KP, Brown JC. Delayed time series predictions with neural networks. Neurocomputing 1998;18(1-3):81–9. https://doi.org/10.1016/S0925-2312(97)00070-2.
[61] Perng CS, Wang H, Zhang SR, Parker DS. Landmarks: a new model for similarity-based pattern querying in time series databases. In: In: Proceedings of 16th International Conference on Data Engineering; 2000. https://doi.org/10.1109/ICDE.2000.839385.
[62] Frank J, Mannor S, Pineau J, Precup D. Time series analysis using geometric template matching. IEEE Trans Pattern Anal Mach Intell 2013;35(3):740–54. https://doi.org/10.1109/TPAMI.2012.121.
[63] Frías-Paredes L, Mallor F, Gastón-Romeo M, León T. Assessing energy forecasting inaccuracy by simultaneously considering temporal and absolute errors. Energy Convers Manag 2017;142:533–46. https://doi.org/10.1016/j.enconman.2017.03.056.
[64] Vallance L, Charbonnier B, Paul N, Dubost S, Blanc P. Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric. Sol Energy 2017;150:408–22. https://doi.org/10.1016/j.solener.2017.04.064.
[65] XGBoost. XGBoos Documentation. https://xgboost.readthedocs.io/en/latest/ [accessed 8 April 2021].
[66] Paraschiv F, Erni D, Pietsch R. The impact of renewable energies on EEX day-ahead electricity prices. Energy Policy 2014;73:196–210.
[67] Erni D. Day-Ahead Electricity Spot Prices - Fundamental Modelling and the Role of Expected Wind Electricity Infeed at the European Energy Exchange. Gallen: University of St.; 2012.
[68] Power trade Skakerag Kraft, https://www.skagerakkraft.no/krafthandel_2/ ; [accessed 28th March 2022].
[69] Wind Power Skakerag Kraft, https://www.statkraft.com/what-we-do/wind-power/ ; [accessed 28th March 2022].
[70] Shao J, Hu K, Wang C, Xue X, Raj B. Is normalization indispensable for training deep neural network?. In: Advances in Neural Information Processing Systems. 2020;vol:33..
[71] Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak 2016;16(3). https://doi.org/10.1186/s12911-016-0318-z.
[72] Acuña E, Rodriguez C. In: Classification, Clustering, and Data Mining Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 639–47. https://doi.org/10.1007/978-3-642-17103-1_60.
[73] Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons; 2004.
[74] Dyer SA, Dyer JS. Cubic-spline interpolation. 1. IEEE Instrum Meas Mag 2001; vol: 4 (2). https://doi.org/10.1109/5289.930984.