A Reinforcement Learning Approach for Dynamic Pricing

Hamid Nakhaei, Shahram Shadrokh
Industrial Engineering Department, Sharif University of Technology, Tehran, Iran

1hamidnakhaei@gmail.com, shadrokh@sharif.edu

Population growth and the increased adoption of electronic devices have led to a surge in electricity demand. However, the construction of adequate infrastructure to meet this additional demand is highly expensive. To address this issue that sometimes causes power outages, dynamic pricing for electricity, a demand response (DR) technique, has gained significant importance. By employing this method, prices can be adjusted on an hourly basis, enabling end consumers (ECs) to regulate their consumption in response to these price fluctuations. The prices can be determined based on different objective functions. This research aims to develop a comprehensive framework for determining the optimal hourly electricity retail prices (HERPs). The primary objectives of this framework are to minimize ECs' costs on electricity and maximize the retailer's profit. First, the wholesale electricity price for each day of the year is forecasted using historical data and the SARIMA method; the selection of the model is such that the residuals are as close to the white noise as possible. Prior to forecasting, the non-stationarity of the data was fixed. Subsequently, the Q-learning algorithm is employed to determine the optimal HERPs for the following day. This process is then applied to Iran's electricity data in the fiscal year 2021-2022 as a case study, and the results are analyzed. Additionally, a sensitivity analysis is conducted on the weighting factor for the retailer's profit and ECs' costs. The findings demonstrate the efficacy of our pricing framework, yielding a notable reduction of 17.1% in annual peak load and 19.4% in total consumption.

Key words: Dynamic Pricing ; Reinforcement Learning ; Demand Response ; Non-stationarity ; Forecasting

---

Nomenclature

Definitions

S_t	random variable of state at time t
n	number of irritations
T	final time step
ε	error term in time series modeling
B	back shift operator
R_t	random variable of reward at time t
A_t	random variable of action at time t
s', s	realizations of state
r	a realization of reward
a	a realization of action
G_t	future discounted return after time step t
q, Q	action-value function
π	policy

Variables

$\widehat{D}_t$	forecasted electricity demand at time step t
D_t	actual electricity demand at time step t

$\widehat{d_t}$	forecasted electricity consumption at time step t
d_t	actual electricity consumption at time step t
$\widehat{\omega}_t$	forecasted electricity wholesale price at time step t
ω_t	actual electricity wholesale price at time step t
rp_t	HERP at time step t
D_t	actual electricity demand at time step t
z	objective function
$\widehat{prof}_t$	forecasted retailer's profit at time step t
$prof_t$	actual retailer's profit at time step t
$\widehat{C}_t$	forecasted ECs' electricity cost at time step t
C_t	actual ECs' electricity cost at time step t
ϕ	coefficients of autoregressive terms
θ	coefficients of moving average terms
δ	average of time series
L	Non-critical load percentage

Parameters

ϵ	epsilon in the epsilon-greedy policy
γ	discount factor in Q-learning
ξ	price elasticity of electricity demand
ρ	weighting factor
α	step size
k_1	coefficient of retailer lowest price
k_2	coefficient of retailer highest price
$\mathfrak{p}, \mathfrak{P}$	orders of autoregressive parts
$\mathfrak{q}, \mathfrak{Q}$	orders of moving average parts
$\mathfrak{d}, \mathfrak{D}$	lags of differencing
$\mathfrak{s}$	seasonality

Others

HERP	hourly electricity retail price
EC	end consumer
ES	energy supplier
DR	demand response
MDP	Markov decision process
RL	reinforcement learning
SARIMA	seasonal autoregressive integrated moving average

1. Introduction

Electricity demand is anticipated to increase by 30% by 2040 compared to 2017 due to an increase in population, urbanization, and the number of electronic devices in use, overloading the existing power grid infrastructure [1]. This gives rise to the importance of demand-side management (DSM) and DR. DR tries to shift consumption away from peak periods when grid load and costs reach their highest points [2]. Dynamic pricing is one of the practices in DR that can influence energy consumption. Dynamic pricing is an approach to determining the optimum prices for products or services in an environment where prices can easily and frequently be altered [3]. The purpose of pricing and revenue optimization is to find the appropriate prices for all products or services, to all customers, through all sale channels at all times [4].

In the context of energy management, dynamic pricing can be used to incentivize ECs to shift their energy consumption from peak to off-peak periods, fostering a flatter load profile. In addition, it can lead to a reduction in ECs' costs of electricity by encouraging consumption when rates are

lower. All the facilities that are specifically allocated to meet peak load, which are large capital investments, stay idle during off-peak periods, resulting in a huge opportunity cost. Despite electricity demand typically being price- and income-inelastic, studies indicate that ECs are responsive to higher prices [5]. Dynamic pricing pilots have demonstrated peak load reductions of up to 58% [6].

Iran has been suffering from a shortage of electricity supply in recent years, leading to prolonged power outages, particularly during warm seasons. Considering the high cost and lengthy timeline of constructing new electricity power plants, adopting dynamic pricing practices becomes increasingly relevant. The fact that electricity retailers in Iran are operating at a loss further underscores the importance of utilizing dynamic pricing in order to increase their profits in this setting.

In this research, we aim to design a framework to determine the optimum HERPs in order to simultaneously maximize the retailer's profit and minimize ECs' costs. The retailer's profit, ECs' costs, and reduction of peak load are analyzed by applying the proposed framework to Iran's electrical power grid data.

This paper is structured as follows: Subsection 1.1 reviews the relative literature on the topic and indicates the existing gap in the literature and the contribution of this study. Section 2 describes our MDP corresponding to the settings of the problem. Section 3 forecasts hourly electricity wholesale prices using a seasonal autoregressive integrated moving average (SARIMA) model. Section 4 implements the Q-learning method to all the data obtained or calculated to find the optimum HERPs. Section 5 discusses the result and conducts a sensitivity analysis, and Section 6 provides the conclusion.

1.1. Literature Review

DSM includes all technologies, activities, and programs on the demand side aimed at reducing energy consumption. Its goal is to lower the overall costs of energy systems, to support policy goals such as emission reduction, and to balance supply and demand. Based on the existing literature, DSM can be classified into three categories: on-site back-up, energy efficiency, and DR. DR uses incentive-based or price-based programs to shift energy consumption from peak to off-peak periods, or more broadly, to change the consumption behavior of ECs [7,8]. The United States Department of Energy defines DR as "a tariff or program established to motivate changes in electric use by end-use customers, in response to changes in the price of electricity over time, or to give incentive payments designed to induce lower electricity use at times of high market prices or when grid reliability is jeopardized" [9]. Reducing or shifting the peak load is more cost-effective than expanding infrastructure to meet the peak load [10]. As an effective method of DSM, DR can either shift or reduce some of the energy consumption during the peak period [11]. Studies have shown that ECs are more willing to reduce rather than reschedule their consumption as a response to dynamic

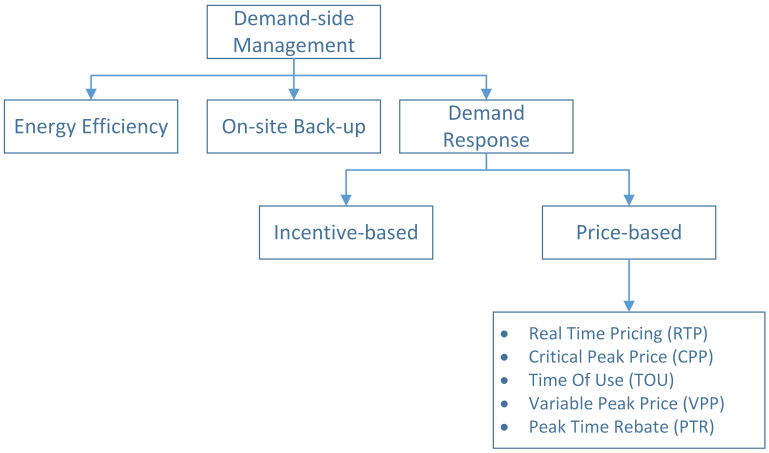


Figure 1 The classification of DSM

pricing. The highest level of DR can be achieved in warm weather and among high-consumption ECs. Additionally, advanced technologies play a significant role in implementing dynamic pricing in the electricity market [5]. The authors in [6] discovered that the presence of enabling technologies, such as in-home displays and programmable or communicating thermostats, can further increase the reduction of peak load. Although, to the date of writing this paper, no enabling technologies exist in Iran's electrical power grid to enable dynamic pricing implementation, this study assumes that dynamic pricing can be implemented to explore potential outcomes.

Price-based DR has been gaining more and more attention in electricity markets in recent years. The presence of a more transparent and competitive wholesale market, developments in communications and enabling technologies, and a growing interest in promoting an advanced and automated power grid among policymakers have contributed to increased interest in price-based DR [12]. There are different price-based DR programs, some of which are regarded as static and others as dynamic [5, 6, 11, 13]. In static strategies such as flat rate tariffs and block rate tariffs, prices do not change with demand changes. On the other hand, in dynamic strategies such as time-of-use (TOU), super peak TOU, critical peak pricing (CPP), variable peak pricing (VPP), real-time pricing (RTP), and peak time rebates (PTR), prices change correspondingly as demand changes. In this context, RTP has the same meaning as dynamic pricing has in pricing and revenue management terminology. For a comprehensive overview, Fig. 1 illustrates the categories of DSM, including the position of RTP within this classification.

A number of studies focused on RTP in power grids to find the optimal pricing policy by considering the retailer's side, the ECs' side, or both sides [14–19]. In [14], an algorithm was developed to compute optimal day-ahead prices, aiming to minimize the retailer's costs while accounting for ECs' willingness to shift their consumption. In [15], the authors presented a day-ahead pricing scheme to maximize social welfare, which is defined as the summation of the retailer's

profit and all ECs' utility functions. The simulation result indicated economic benefits for both retailers and ECs. Similar results were obtained in [16], where the authors proposed an RTP framework using an objective function to maximize the retailer's profit while ensuring that the ECs' costs do not exceed pre-implementation status. In [17], an RTP program was proposed by minimizing retailer costs and incorporating self- and cross-time elasticity factors. The authors in [18] studied a setting in which the retailer first clusters ECs and determines the optimum prices for each cluster so as to maximize its profit. Subsequently, ECs reschedule their consumption in response to the prices in order to minimize their costs. Game theory approaches are also used to model and solve RTP [19]. However, all of these papers consider the electricity market deterministic, whereas this environment is totally dynamic.

All of the above-mentioned studies suffer from two shortcomings. First, they ignored the presence of uncertainty, which is a crucial aspect in electricity markets. Disregarding uncertainty may cause the results to deviate from being optimal and impose costs on both ECs and retailers. Both electricity demand and wholesale prices involve uncertainty to some extent due to the volatile nature of the market. Therefore, it is of paramount importance to incorporate uncertainty into the process of modeling as well as solving electricity market dynamic pricing. Second, these studies did not take advantage of modeling dynamic pricing as a sequential decision-making problem, as it can better reflect the essence of the problem since dynamic pricing is about finding optimal prices in a sequential manner. In fact, artificial intelligence (AI) tools can be used alongside modeling dynamic pricing as a sequential decision-making problem to tackle the complexities of this problem.

The rapid growth of the Internet has contributed to rich data sets of purchase histories by consumers across all industries. Moreover, the advancement in technologies has enabled researchers and scholars to adopt more complex and advanced methods, such as AI, to address bigger problems. Recent studies have focused more on AI applications in the domain of DR [20]. Reinforcement learning (RL), one of the paradigms of AI, has been gaining increasing attention in dynamic pricing, as it can benefit from both the sequential nature of modeling problems as a Markov decision process (MDP) and the computational potential of AI. RL is a learning algorithm to find the best map from states to action by maximizing a cumulative reward. To be more specific, the agent interacts with the environment and, through trial-and-error, learns the best action to take in a specific state. The interaction between the agent and the environment is illustrated in Fig. 2.

The studies in [21–30] adopt RL approaches to model dynamic pricing DR in power grids and obtain optimal prices within the stochastic environment of the electricity market. The authors of [21] developed both feedforward and recursive neural networks to model electricity demand and proposed DR pricing based on this modeling. However, this approach primarily focused on minimizing the costs of ECs. In [22], the selection between various retailers with different pricing

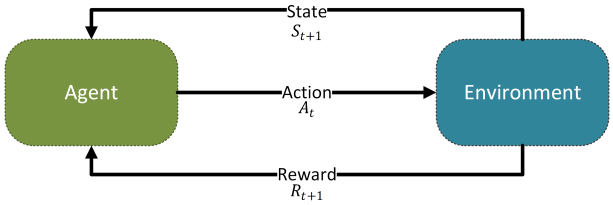


Figure 2 The agent-environment interaction

policies from the ECs' perspective was investigated. A Q-learning algorithm was designed to solve the problem with the purpose of minimizing ECs' costs. In [23], DR dynamic pricing was modeled as an MDP aiming at minimizing ECs' costs and maximizing the retailer's profit. In this study, ECs' demand was assumed to follow a Poisson distribution. To solve the problem, a Q-learning method was developed, considering computational complexity and convergence speed. The same authors continued this study in [24], where ECs' energy consumption scheduling was included in the system. They extended their previous study by using a multi-agent learning structure in which the retailer decides on the optimum prices, and, after that, ECs decide on their consumption schedule to further minimize their cost. Despite their advancements, these two papers crucially depend on state transition probabilities, which may not be easily available in the real world. Moreover, considering the demand as a Poisson distribution cannot reflect its nonstationary nature. It is noteworthy that both demand and wholesale prices are extremely nonstationary and should be taken care of. In [25], an electricity market model was presented, and a Q-learning algorithm was introduced to solve the dynamic pricing problem. This work continued in [26], where the objective function was defined as a weighted summation of ECs' costs and the retailer's profit over a period of time. Then, a tabular Q-learning algorithm was used to find the best pricing policy, as the state space and the action space were chosen to be discrete. The simulation results showed the potential of the proposed dynamic pricing for reducing ECs' costs and promoting the retailer's profit. The studies in [27, 28] share the same model and fundamental principles. They both used an LSTM model to forecast demand and wholesale prices and used the results as inputs to the Q-learning algorithm. However, in [28], plug-in electric vehicles were incorporated into the model. The authors of [29] formulated DR dynamic pricing using MDP and employed RL to solve it. The results indicated that ECs' costs decreased as a result of either reducing or shifting energy consumption. In [30], two independent problems were modeled by the MDP framework, one of which aimed at maximizing ECs' welfare and the other at the retailer's profit. A distributed algorithm using RL was developed to solve the problem.

Despite the extensive research conducted on dynamic pricing in electricity markets, there are several limitations within the aforementioned studies. First and foremost, to the best of our knowledge no study on dynamic pricing in power grids has adequately addressed the nonstationary

nature of this market. The majority of studies simply used historical data or a probability distribution as demand in their model evaluation, whereas in the real world, knowledge of demand and wholesale prices is not available a priori. Although a few studies have attempted to forecast demand and wholesale prices and included these forecasts in their models, they failed to account for non-stationarity. Especially when it comes to long-run demand and wholesale price forecasting, non-stationarity significantly impacts the reliability of the forecast. Therefore, our study defers from the literature by proposing a pricing framework that can be relied upon in the long run. Secondly, there is no study on dynamic pricing in Iran's power grid using new methodologies such as RL. Considering loss-making retailers in Iran and frequent power outages, it is of paramount importance to take a closer look at this market and find a solution to these problems. This is the first study that develops an RL algorithm for dynamic pricing on Iran's power grid real data.

In this study, several steps are taken to address these challenges. Firstly, we construct an electricity market representation corresponding to the real conditions in Iran's power grid. We adopt the SARIMA method to forecast day-ahead wholesale prices and demand. Employing SARIMA to forecast wholesale prices fundamentally enables us to capture and mitigate non-stationarity. Subsequently, we model the problem as an MDP, integrating the forecasted wholesale prices and demand as components of the MDP. The objective function is to find the optimum HERPs, such that simultaneously maximize the retailer's profit and minimize ECs' costs. Finally, a Q-learning algorithm, which is model-free and independent of transition probabilities, is developed to find the solutions. Hence, the main contributions of this study can be summarized as follows:

- In an attempt to get closer to reality, we integrate demand and wholesale price forecasting into the RL methodology. The demand and wholesale prices are extremely nonstationary as they involve multi-seasonality, trends, and volatility. We use SARIMA as a time series method to deal with these challenges and develop a dynamic pricing framework that can be relied upon in the long run.
- Given the significant number of ECs and potential multiple energy suppliers (ESs) within a power grid, processing individual demand and wholesale price data is prohibitively costly. Therefore, we use aggregate data for both demand and wholesale prices in our analysis. This approach reduces computational expenses and eliminates the need for additional infrastructure.
- Iran's power grid has been suffering from a supply-and-demand imbalance in recent years, while retailers in this market are operating at a considerable loss. To the best of our knowledge, our study stands as first to apply RL methods to the real data and model of Iran's electricity market.

2. Modeling

Iran's electricity market consists of both wholesale and retail markets, involving three key stakeholders: ESs, retailers, and ECs. In the wholesale market, multiple ESs generate electricity, which

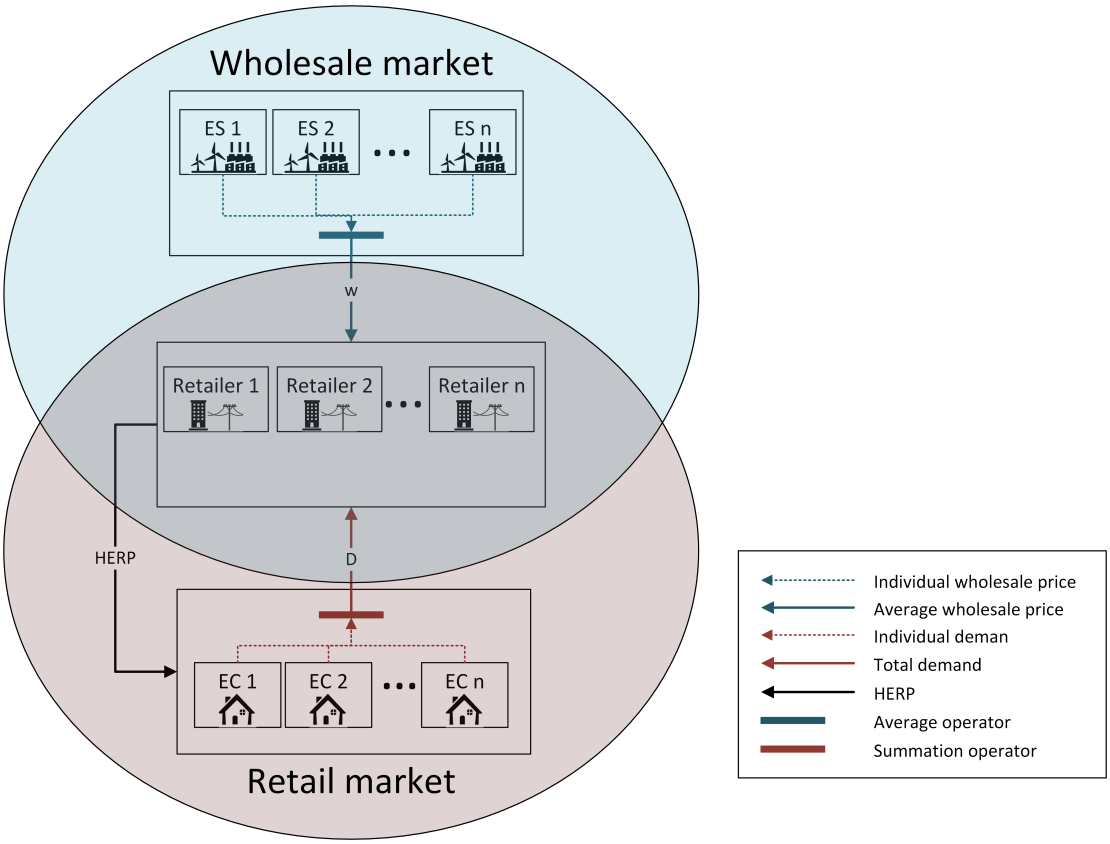


Figure 3 The hierarchy of electricity market

is then sold to retailers. Subsequently, in the retail market, retailers distribute and sell electricity to ECs with the aim of making a profit. For the sake of simplicity and efficiency, it is assumed that there is only one ES and one retailer, since using aggregate data for both demand and wholesale prices leads to less processing and needed computational resources. Under this assumption, all consumers across various sectors, such as commercial, household, industrial, and others, are treated as ECs, and their demand is considered collectively. Similarly, wholesale prices from different ESs are not differentiated. Rather, the total demand and average wholesale prices are used. This hierarchy of electricity market is illustrated in Fig. 3. It is assumed that the retailer cannot affect wholesale prices but can decide on the HERPs in order to maximize its profit. Accordingly, the ECs can adapt their consumption behavior in response to to the dynamic pricing policy implemented by the retailer and change their consumption pattern to minimize their electricity costs.

To model the dynamic described above as a discrete finite horizon MDP, all the corresponding components are defined. In the settings of this study, at each time step (one hour of each day), the retailer (agent) knows the forecasted electricity demand and wholesale price (state). It chooses the HERP (action) for that time step and informs the environment (ECs and the ES) of its action. ECs adapt their electricity demand based on the price and their price elasticity of electricity demand,

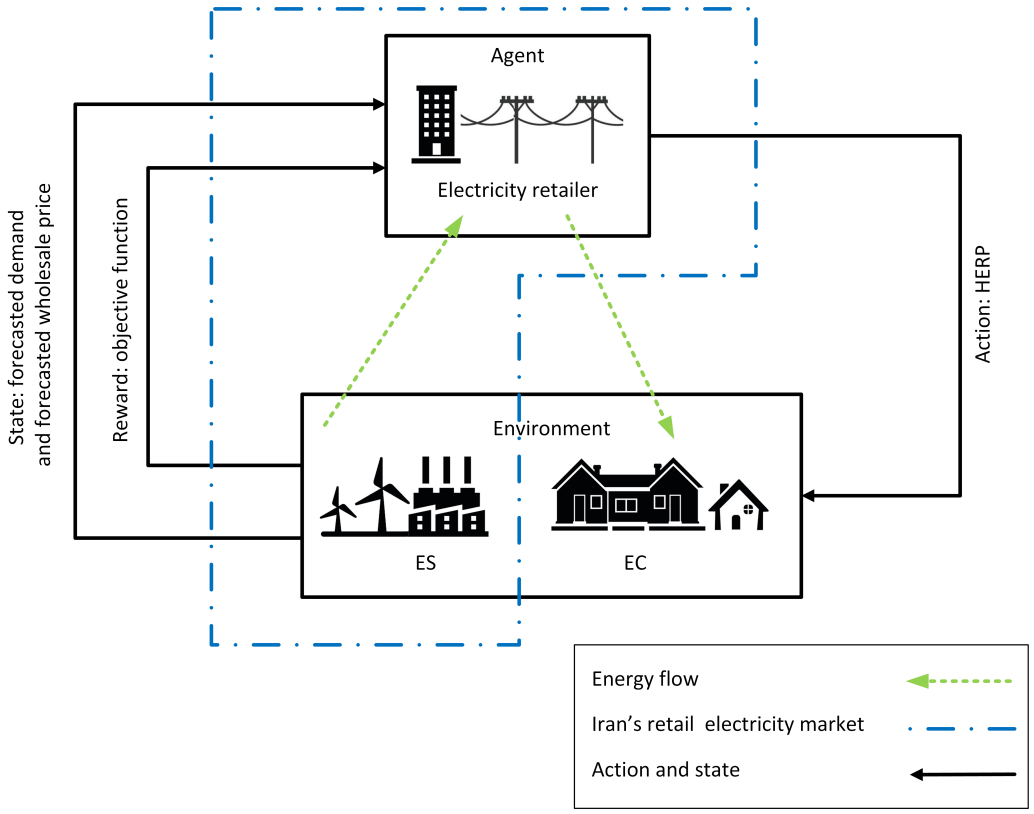


Figure 4 **The MDP model of the electricity market**

and according to the actual electricity consumption, a reward will be calculated to maximize the retailer's profit and minimize ECs' electricity costs. Then, the system moves on to the next time step. Thus, there is a sequence like this:

$$S_t\left(\widehat{D}_t,\widehat{\omega}_t\right) \rightarrow rp_t \rightarrow R_t \rightarrow S_{t+1}\left(\widehat{D}_{t+1},\widehat{\omega}_{t+1}\right) \rightarrow rp_{t+1} \rightarrow R_{t+1} \rightarrow \cdots \tag{1}$$

In the real world, ECs decide on the actual consumption only after HERPs are available. On the other hand, the retailer can determine HERPs only after they know how much the consumption and the cost of buying from the ES will be. To solve this problem, the retailer uses forecasted electricity consumption and wholesale prices instead of the actual ones as its state. Therefore, forecasting of electricity consumption and wholesale prices is needed. The MDP model of the problem is illustrated in Fig. 4.

The concept of price elasticity of demand is used to define actual electricity consumption in terms of actual electricity demand. Price elasticity of demand is defined as the percentage change in demand in response to a one percent change in price. We assume that, before enforcing the new pricing policy, the electricity demand is met at the wholesale price and the retailer does not charge ECs more. Based upon this definition, the mathematical presentation is as follows.

$$\xi = \frac{d_t - D_t}{rp_t - \omega_t} \times \frac{\omega_t}{D_t} \tag{2}$$

$$\xi < 0$$

$$\xi \times (rp_t - \omega_t) \times \frac{D_t}{\omega_t} = d_t - D_t \Rightarrow \tag{3}$$
$$d_t = D_t \times \left(1 + \xi \times \frac{rp_t - \omega_t}{\omega_t}\right)$$

It is worth mentioning that electricity demand can be divided into two parts: non-critical load and critical load. The critical load is a proportion of electricity demand that is indispensable and must be met, such as the electricity needed for refrigerators or lights during nighttime. The non-critical load, on the other hand, refers to the remaining proportion of demand, which can be adjusted in response to HERP changes. Electricity used for heating and air conditioning during peak periods is an example of the non-critical load. We assume that $L\%$ of total demand is the non-critical load. Thus, the following equation holds true:

$$d_t = L\% \times D_t \times \left(1 + \xi \times \frac{rp_t - \omega_t}{\omega_t}\right) + (1 - L)\,\% \times D_t \tag{4}$$

It should be noted that the non-critical load cannot be decreased to a negative amount. Given that ξ is a negative quantity, the HERP can be raised until the non-critical load reaches zero.

$$1 + \xi \times \frac{rp_t - \omega_t}{\omega_t} \geq 0 \Rightarrow \tag{5}$$
$$\xi\omega_t - \xi rp_t \leq \omega_t \Rightarrow$$
$$rp_t \leq \frac{\omega_t\,(\xi - 1)}{\xi}$$

The reward of the MDP should be designed in a way that reflects the purpose of this study: to maximize the retailer's profit and minimize ECs' costs. They are defined as follows:

$$prof_t = (rp_t - \omega_t) \times d_t \tag{6}$$
$$C_t = rp_t \times d_t$$

The retailer's profit at time step t is calculated as the difference between the revenue obtained from end consumers and the cost of purchasing electricity from the wholesale market. The cost for ECs at time step t is computed as the product of the HERP set by the retailer and the actual electricity demand. It is important to note that Eq. (6) use the actual consumption of electricity. However, since the actual quantities for these variables are not available at the time of decision-making, the forecasted quantities should be substituted as follows:

$$\widehat{d_t} = L\% \times \widehat{D}_t \times \left(1 + \xi \times \frac{rp_t - \widehat{\omega}_t}{\widehat{\omega}_t}\right) + (1 - L)\,\% \times \widehat{D} \tag{7}$$
$$rp_t \leq \frac{\widehat{\omega}_t\,(\xi - 1)}{\xi}$$
$$\widehat{prof}_t = (rp_t - \widehat{\omega}_t) \times \widehat{d_t}$$
$$\widehat{C_t} = rp_t \times \widehat{d_t}$$

By using an appropriate weighting factor, the reward and return of the MDP are defined as follows:

$$R_t = \rho\,(rp_t - \widehat{\omega}_t) \times \widehat{d}_t - (1-\rho)\,rp_t \times \widehat{d}_t; \qquad 0 \le \rho \le 1 \tag{8}$$
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{T-t} \gamma^k R_{t+k+1}; \qquad 0 \le \gamma \le 1$$

Accordingly, the objective function that we seek to optimize is as follows:

$$z = \max \sum_{t=1}^{T} \left(\rho\,(rp_t - \widehat{\omega}_t) \times \widehat{d}_t - (1-\rho)\,rp_t \times \widehat{d}_t \right) \tag{9}$$
$$\widehat{\omega}_t \le rp_t \le \frac{\widehat{\omega}_t\,(\xi - 1)}{\xi} \tag{10}$$
$$k_1 \min_t \omega_t \le rp_t \le k_2 \max_t \omega_t \tag{11}$$

The first constraint ensures that, at any given time step, the HERP will never be lower than the wholesale price, and it will never be so high that the non-critical load becomes negative. The second constraint sets a boundary for the HERP and helps define a set of discrete prices from which the HERP for each time step is chosen.

All the data used in this study was obtained from Iran Grid Management Company (IGMC), available at [31]. As stated earlier, a forecast of electricity demand and wholesale prices is needed. Fortunately, IGMC provides us with a reliable hourly forecast of electricity demand for the next day, issued daily at 10 a.m., with an $R^2 = 0.98$, which is completely acceptable. Nevertheless, there is no forecast available for hourly wholesale prices. Thus, each day, historical data is used to forecast hourly wholesale prices for the following day at 11 a.m. This ensures that all necessary information for executing the Q-learning algorithm is available by 11 a.m. Being the most up-to-date data available at the time of writing this study, the full-year interval from March 21$^{\text{st}}$, 2021 (the first day of the year in the Iranian calendar) to March 21$^{\text{st}}$, 2022, is chosen to implement the proposed model and analyze results and the behavior of the model in the long run.

Due to the lack of data on the critical and non-critical loads in Iran, we rely on estimation. Based on Iran's electricity tariffs [32], the minimum necessary electric power for households, which can be considered the critical load, during 2021-2022 was 200 kW per month. Moreover, statistics indicate that approximately 32% of total ECs belong to the household sector, with a count of 30885000 household ECs during the same time period [33]. However, to extrapolate these estimates to other sectors like industrial and agricultural, further assumptions are required. For the sake of simplicity, we assume uniform critical load requirements across all sectors, although this assumption may

oversimplify the actual scenario. With these considerations in mind, the hourly total critical load can be estimated as follows:

$$\frac{0.2\text{MWh}}{30\times 24}\,(\text{hourly critical laod of each household})\times 30885000\,(\text{number of households})\times\frac{1}{0.32}=26715.09\text{MWh} \quad (12)$$

Dividing the hourly total critical load (26715 MWh) by the hourly total average consumption in the same period (33647 MWh) results in an approximate value of L=20%.

For the rest of this study, the unit of forecasted and actual electricity demand is MWh, and the unit of price is $^{\text{IRR}[1]}/_{\text{kWh}}$.

3. Forecasting

Forecasting prices has gained incredible attention, especially after the emergence of neural networks. Insights gained from the forecasted prices have numerous applications both in academia and industry, particularly when the accuracy and reliability of the forecasting model are satisfactory. Among all methods of forecasting, we opt to use the SARIMA method, as it can take the non-stationarity nature of prices into account. Hence, its forecasting can be relied upon in the long run. Fig. 5 shows the time series plot of ω_t for both the 2020-2022 period and a ten-day interval. As shown in Fig. 5, the wholesale prices have both trend and seasonal components.

The wholesale prices for the following day are forecasted every day at 10 a.m., utilizing the most recent four months of data. Therefore, the forecast horizon or lead time is 38. The forecast interval is an hour, and we employed a rolling horizon forecasting approach.

In time-series methodologies, an $ARIMA(\mathfrak{p},\mathfrak{d},\mathfrak{q})$ process can be presented as below:

$$X_t=(1-B)^{\mathfrak{d}}\,y_t \quad (13)$$

$$X_t-\phi_1X_{t-1}-\cdots-\phi_{\mathfrak{p}}X_{t-\mathfrak{p}}=\delta+\epsilon_t-\theta_1\epsilon_{t-1}+\cdots+\theta_{\mathfrak{q}}\epsilon_{t-\mathfrak{q}}$$

$$X_t=\delta+\sum_{i=1}^{\mathfrak{p}}\phi_iX_{t-i}+\epsilon_t-\sum_{i=1}^{\mathfrak{q}}\theta_i\epsilon_{t-i}$$

The equation can be rewritten using the backshift operator as follows:

$$\Phi(B)(1-B)^{\mathfrak{d}}\,y_t=\delta+\Theta(B)\,\epsilon_t \quad (14)$$

In addition, a $SARIMA(\mathfrak{p},\mathfrak{d},\mathfrak{q})\times(\mathfrak{P},\mathfrak{D},\mathfrak{Q})_{\mathfrak{s}}$ model is formulated as below:

$$\Phi^*(B^{\mathfrak{s}})\,\Phi(B)(1-B)^{\mathfrak{d}}(1-B^{\mathfrak{s}})^{\mathfrak{D}}\,y_t=\delta+\Theta^*(B^{\mathfrak{s}})\,\Theta(B)\,\epsilon_t \quad (15)$$

[1] Iranian rial

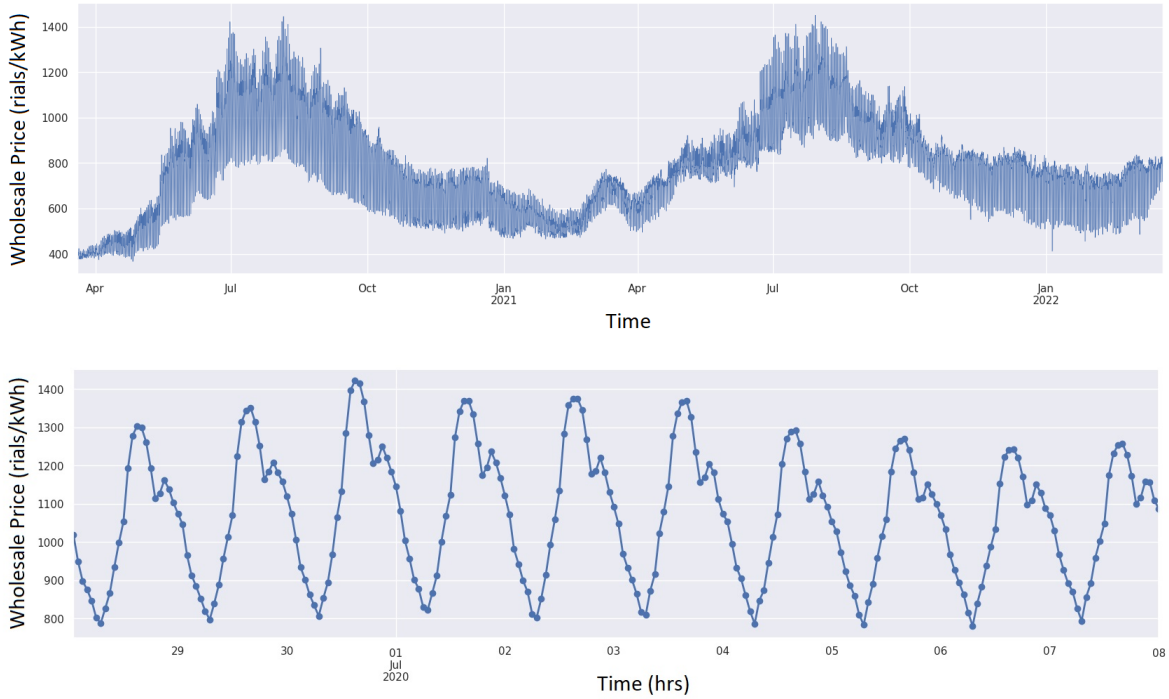


Figure 5 Time series plot of hourly wholesale price

In ARIMA and SARIMA models, it is assumed that ϵ_t is white noise.

The prerequisite to implementing the SARIMA model on a data set is that the data set be stationary. As the time series plot in Fig. 5 demonstrates, the data has a 24-hour seasonality and a trend, and the standard deviation of the data is not constant over time, making the time series nonstationary. Moreover, if a time series is stationary, its ACF and PACF either decay exponentially or cut off after a certain lag. It is not the case, as shown by the ACF and PACF plots of the data in Fig. 6.

To overcome the trend and the 24-hour seasonality, a 1-lag differencing and a 24-lag seasonal differencing are performed, respectively. A $\mathfrak{d}$-lag differencing is given in Eq. (16) [34].

$$x_t = y_t - y_{t-\mathfrak{d}} = y_t - B^{\mathfrak{d}} y_t = \left(1 - B^{\mathfrak{d}}\right) y_t \tag{16}$$

Although the ACF and PACF plots of the data after performing the mentioned differencings, shown in Fig. 7, indicate a stationary time series, it seems that the dependence of the standard deviation on time persists. To mitigate this effect, a Box Cox transformation, which is a power transformation, is applied to the time series.

To ensure the stationarity of the time series data, an augmented Dickey-Fuller test is performed on the differenced data. The test results, summarized in Table 1, reveal a small p-value. Therefore, the null hypothesis, or non-stationarity of the data, is rejected, and we can assume that the time series is stationary and ready to implement the SARIMA model.

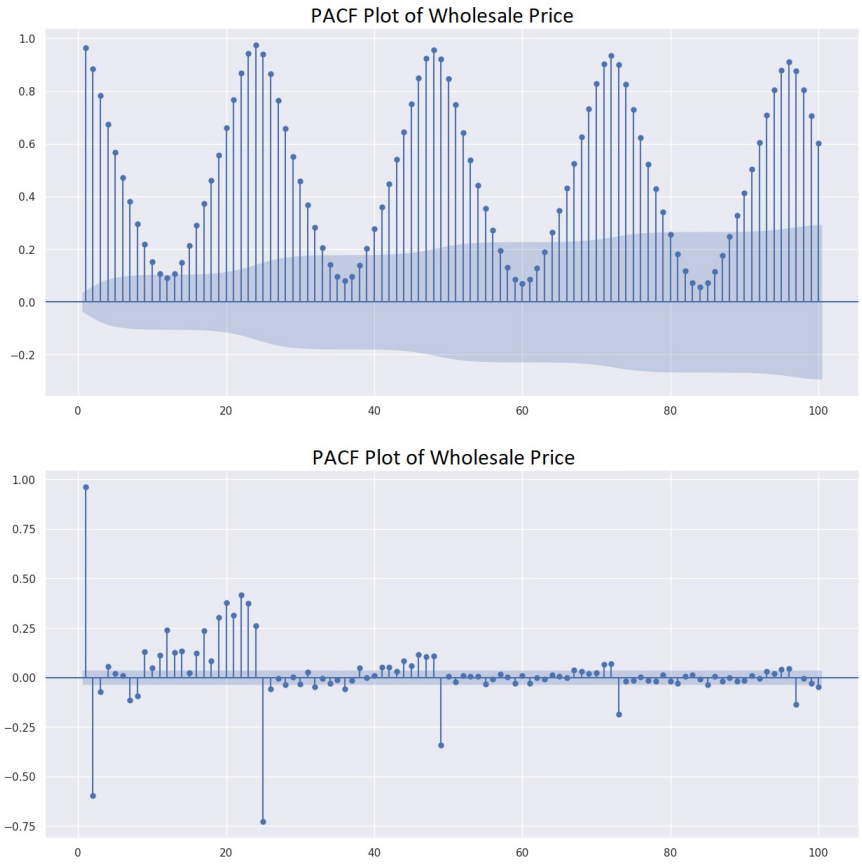


Figure 6 **The ACF and PACF plot of wholesale prices**

Table 1 **Augmented Dickey-Fuller test results**

Test statistic	-15.087
P-value	8.29e-28
Sample size	2815
Critical value at 1% significance level	-3.432
Critical value at 5% significance level	-2.862
Critical value at 10% significance level	-2.567

Model identification plays a crucial role in implementing the SARIMA method, as it involves determining the order of different components of the model. This process often begins with analyzing ACF and PACF plots. As shown in Fig. 7, the seasonal part of the ACF plot cuts off after the first lag, while the PACF plot decays exponentially, meaning it follows an $MA(1)$ process. Therefore, we establish the seasonal model order as $(\mathfrak{P}, 1, \mathfrak{Q})_{24} = (0, 1, 1)_{24}$.

To identify the order of the regular (non-seasonal) part of the model, an iterative approach is employed. In this approach, residuals are analyzed to make sure they converge to white noise with no significant autocorrelation or partial autocorrelation values (i.e., the ACF and PACF should not differ significantly from zero for all lags greater than 1). Through this iterative process, the order of

Figure 7 Time series, ACF, and PACF plots after differencing

the model is identified as $(\mathfrak{p},\mathfrak{d},\mathfrak{q}) \times (\mathfrak{P},\mathfrak{D},\mathfrak{Q})_{24} = (0,1,[1,2,3,5,7,8]) \times (0,1,1)_{24}$. The parameters are estimated using Python and summarized in Table 2.

The model can be written as Eq. (17).

$$(1-B)\left(1-B^{24}\right)\omega_t =$$

Table 2 Estimated parameter of the SARIMA model

	Estimate	Std. Error	Z value	Pr(<| z |)
ma.L1	-0.1074	0.017	-6.229	0.000
ma.L2	-0.1428	0.016	-9.078	0.000
ma.L3	-0.1093	0.015	-7.531	0.000
ma.L5	-0.0911	0.018	-5.099	0.000
ma.L7	-0.0766	0.017	-4.595	0.000
ma.L8	-0.0533	0.017	-3.106	0.002
ma.S.L24	-0.6068	0.011	-55.459	0.000
sigma2	0.0269	0.000	66.985	0.000
AIC = -2183.717 ; BIC = -2136.102				

Table 3 Forecasting Model Performance Metrics

R^2	0.946
MSE	1559.8
MAE	27.4

$$\left(1-0.1074B-0.1428B^2-0.1093B^3-0.0911B^5-0.0766B^7-0.0533B^8\right)\left(1-0.6068B^{24}\right)\epsilon_t \quad (17)$$

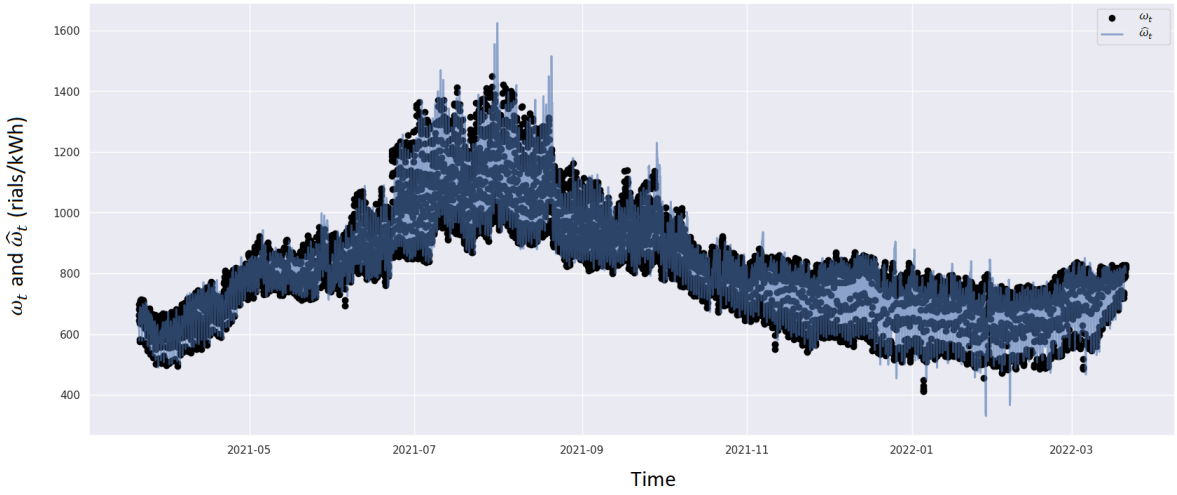
The time series plots of actual and forecasted wholesale prices for the entire 2021-2022 year and a random interval of 10 days are illustrated in Fig. 8. Additionally, Table 3 summarizes key metrics including R-squared, mean squared error (MSE), and mean absolute error (MAE). These metrics provide valuable insights into the accuracy and performance of the forecasting model.

It is expected that the residuals be normally and independently distributed with mean 0, and variance σ^2 (i.e., $NID\,(0,\sigma^2)$). This means they should follow a normal distribution with mean 0 and variance σ^2, be independent, exhibit no autocorrelation, and maintain a constant variance over time. In this regard, Fig. 9 shows four plots of the residuals for the SARIMA model.

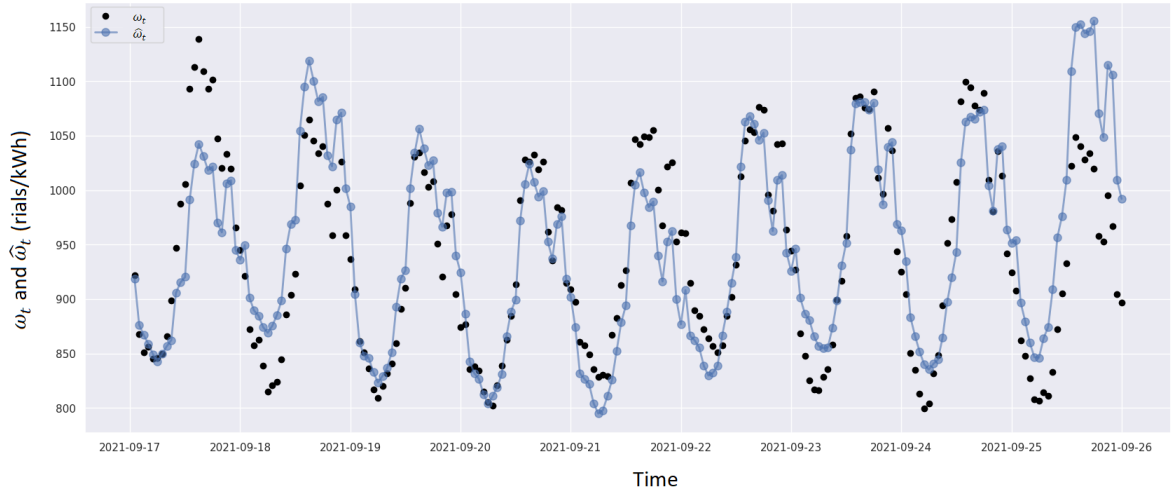
All the $NID\,(0,\sigma^2)$ criteria are seemingly met, except for a few extreme data points, preventing the distribution from being perfectly normal, which are natural to occur in the domain of price forecasting. Furthermore, the Ljung–Box test is performed to make sure no significant autocorrelation in the residuals exists. As the p-value of this test is 0.58, we can assume that the residuals are not correlated.

4. Q-learning

Tabular methods of reinforcement learning algorithms can be fundamentally classified into three main categories: dynamic programming, Monte Carlo, and temporal difference learning methods. Each method has its strengths and weaknesses, making it suitable for specific applications. Dynamic programming methods require the complete model of the environment. This model is presented as the dynamics function $p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0,1]$, defined as $p(s', r \mid s, a) =$



(a) 2021-2022 interval



(b) A random 10-day interval

Figure 8 Actual and Forecasted Wholesale Prices

$\Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$. Since the environment's model is not available in this study, it is not possible to use dynamic programming approaches. However, Monte Carlo and temporal difference methods do not require a model, which is why they are called model-free methods. In the context of reinforcement learning, bootstrapping means estimating the value function based on its previous estimates, which enhances speed and reduces computational resources in need. Since Monte Carlo methods do not bootstrap, temporal difference methods are of interest in this study.

There are mainly two different types of temporal difference methods, namely on-policy and off-policy learning. On-policy learning involves evaluating or improving the policy from which actions are taken, whereas off-policy learning involves taking actions from a behavior policy that differs

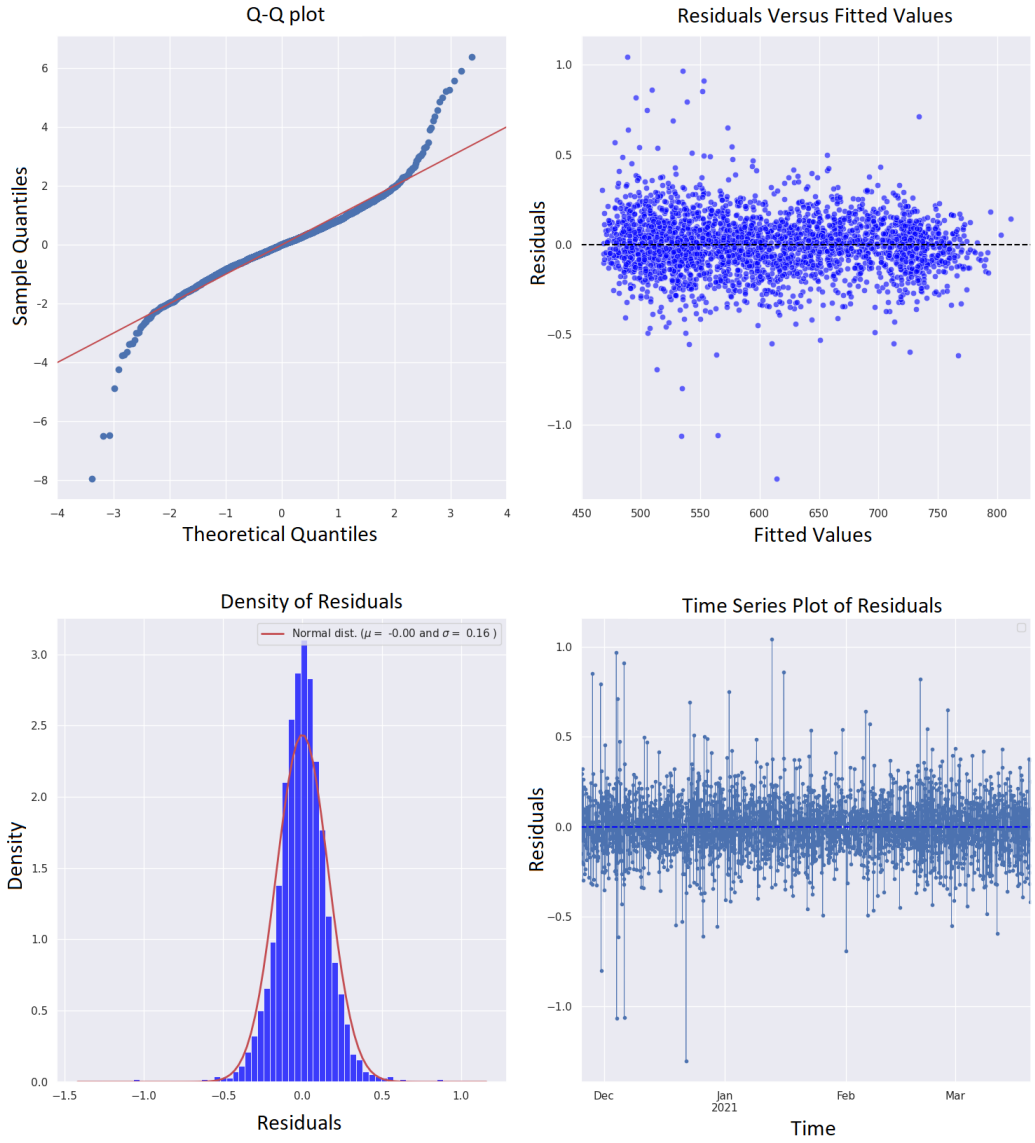


Figure 9 Residual plots for the SARIMA model

from the policy being learned about, also known as the target policy. Off-policy methods are well-suited for situations in which implementing the policy in the environment and receiving feedback are not feasible. In fact, the actions are taken from a different policy, but the target policy is being evaluated and improved. SARSA and Q-learning are two of the most well-known methods of on-policy and off-policy learning, respectively [35]. Due to the nature of pricing problems and their potential risk of imposing losses on both the retailer and ECs, the off-policy learning method, Q-learning, is adopted in this study.

Q-learning has been considered as one of the most significant and remarkable advances in reinforcement learning. In this algorithm, introduced by Watkins [36], the action-value function, Q, is

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
    $S \leftarrow S'$
  until S is terminal

Figure 10 Tabular Q-learning algorithm [35]

updated as below:

$$Q\left(S_t, A_t\right) \leftarrow Q\left(S_t, A_t\right) + \alpha\left[R_{t+1} + \gamma \max_a Q\left(S_{t+1}, A_t\right) - Q\left(S_t, A_t\right)\right] \tag{18}$$

In this case, given some conditions, Q converges to the optimal action-value function, q^*, independent of the behavior policy. In other words, the target policy will converge to the optimal policy, regardless of how the actions are taken. Having the optimal q^*, the optimal policy can be determined easily: in each state, the best action is where q^* is maximum.

$$a^* = \text{argmax } q^*\left(S_t, A_t\left(S_t\right)\right); \qquad \forall S \in \mathcal{S} \tag{19}$$

In tabular Q-learning approaches, the action-value function manifests as a table (i.e., the Q-table) in which each row represents a state, each column represents an action, and each cell represents the action-value (i.e., the expected cumulative future discounted reward) of that state-action pair. This necessitates discrete states and actions. The pseudocode of Q-learning is shown in Fig. 10.

The convergence of Q-learning is guaranteed under the following sufficient conditions [36]:

1. There should be sufficient observations of all state-space pairs (i.e., an infinite number of observations in an infinite number of iterations).

2. The step size should be positive, monotonically decreasing, and converge to zero as the number of iterations approaches infinity.

3. The sum of all the step sizes should converge to infinity as the number of iterations approaches infinity.

To ensure these conditions are met, an epsilon-greedy behavior policy is commonly used. In an epsilon-greedy policy, the probability distribution of action selection is defined as:

$$\Pr\left\{A_t = a_t\right\} = \begin{cases} \frac{\varepsilon}{|\mathcal{A}(S_t)|} & \forall\, a_t \neq \text{argmax } Q\left(S_t, A_t\right) \\ 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(S_t)|} & \forall\, a_t = \text{argmax } Q\left(S_t, A_t\right) \end{cases} \tag{20}$$

Table 4 **Selected value for the model parameters**

Parameter	Value
k_1	1.2
k_2	1.8
ξ	-0.9
γ	0.95
ρ	0.5

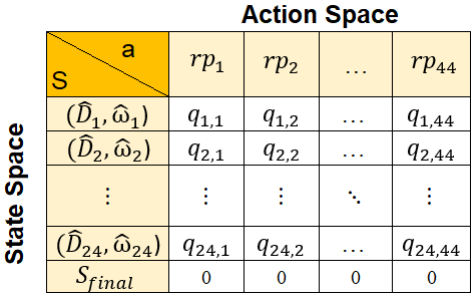


Figure 11 **An illustration of the Q-table**

To satisfy the first sufficient condition, ε is set to 0.05, and to satisfy the last two sufficient conditions, the step size is defined as $\alpha = \frac{1}{1+\lfloor \frac{n}{200} \rfloor}$, where n is the number of iterations, and $\lfloor\ \rfloor$ denotes the floor function.

The next step in the Q-learning algorithm is to define and initialize the Q-table. To accomplish this, feasible actions are first identified based on Eq. (11). Since the maximum and minimum wholesale prices of the next year are not known at the point of decision-making, the equivalent values from the previous year are considered as $\min_t \omega_t$ and $\max_t \omega_t$. Table 4 lists the selected values for k_1, k_2, and other model parameters. Given these parameters, the set of all feasible actions will contain multiples of 50 between 440 and 2590 $^{\text{IRR}}/_{\text{kWh}}$, resulting in a total of 44 actions.

The Q-table has 25 rows (24 rows for 24 states of the day and 1 row for the terminal state) and 44 columns corresponding to the 44 feasible actions, as shown in Fig. 11. According to Eq. (10), not all actions are available in all states. To initialize the Q-table, all cells are assigned a value of 0, except for cells representing infeasible actions, which are assigned negative infinity. The Q-table initialization algorithm is illustrated in Fig. 12.

The price elasticity of demand is a significant parameter in the model and influences results to a great extent. A number of studies have investigated the price elasticity of Iran's electricity demand, suggesting different estimates of this parameter [37–43]. For instance, the authors in [41] estimated the price elasticity of demand to be within the range of -0.87 to -1.02. Taking into account other studies, ξ is set to -0.9. Additionally, the discount factor and weighting factor are assumed to be 0.95 and 0.5, respectively.

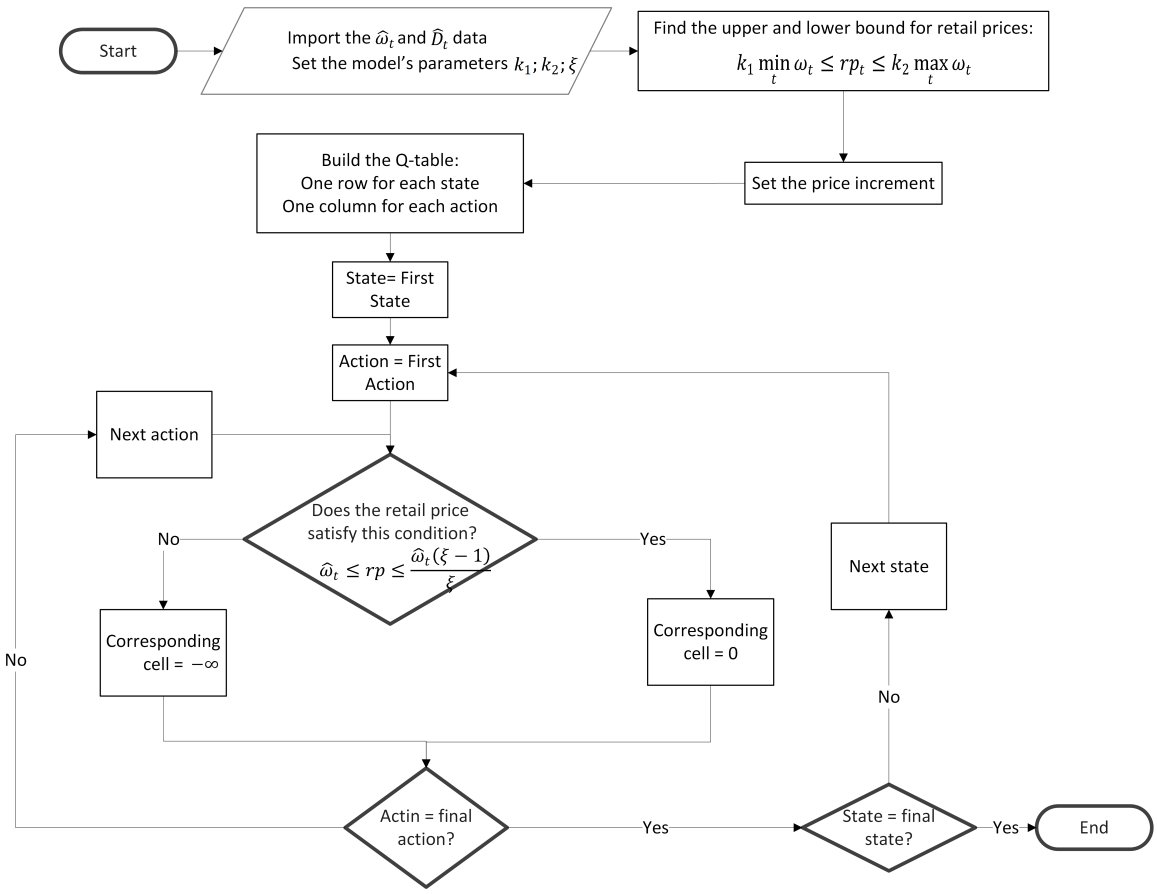


Figure 12 The Q-table initialization algorithm

The proposed Q-learning algorithm is shown in Fig. 13. The algorithm will not stop until the difference between present and previous estimations of action-values becomes less than 1.

To investigate the convergence of the algorithm, the average of action-values and reward on the 100^{th} day as an illustrative example is shown in Fig. 14. The difference between present and previous estimations of action-values converges to zero as the number of iterations increases, as shown in Fig. 15.

Building on the preceding discussions, we propose a framework to solve dynamic pricing in an electricity market, as shown in Fig. 16. In the first phase, forecasting, the wholesale price and, if necessary, demand data are imported into the model. Following the transformation of the data into stationary time series, SARIMA model identification and parameter estimation are implemented. After ensuring that the behavior of residuals is desirable, we proceed to the second phase: learning. In this phase, the inputs consist of forecasted wholesale prices and demand. Once parameters are set and the learning algorithm is implemented, the framework yields its outcome—HERPs.

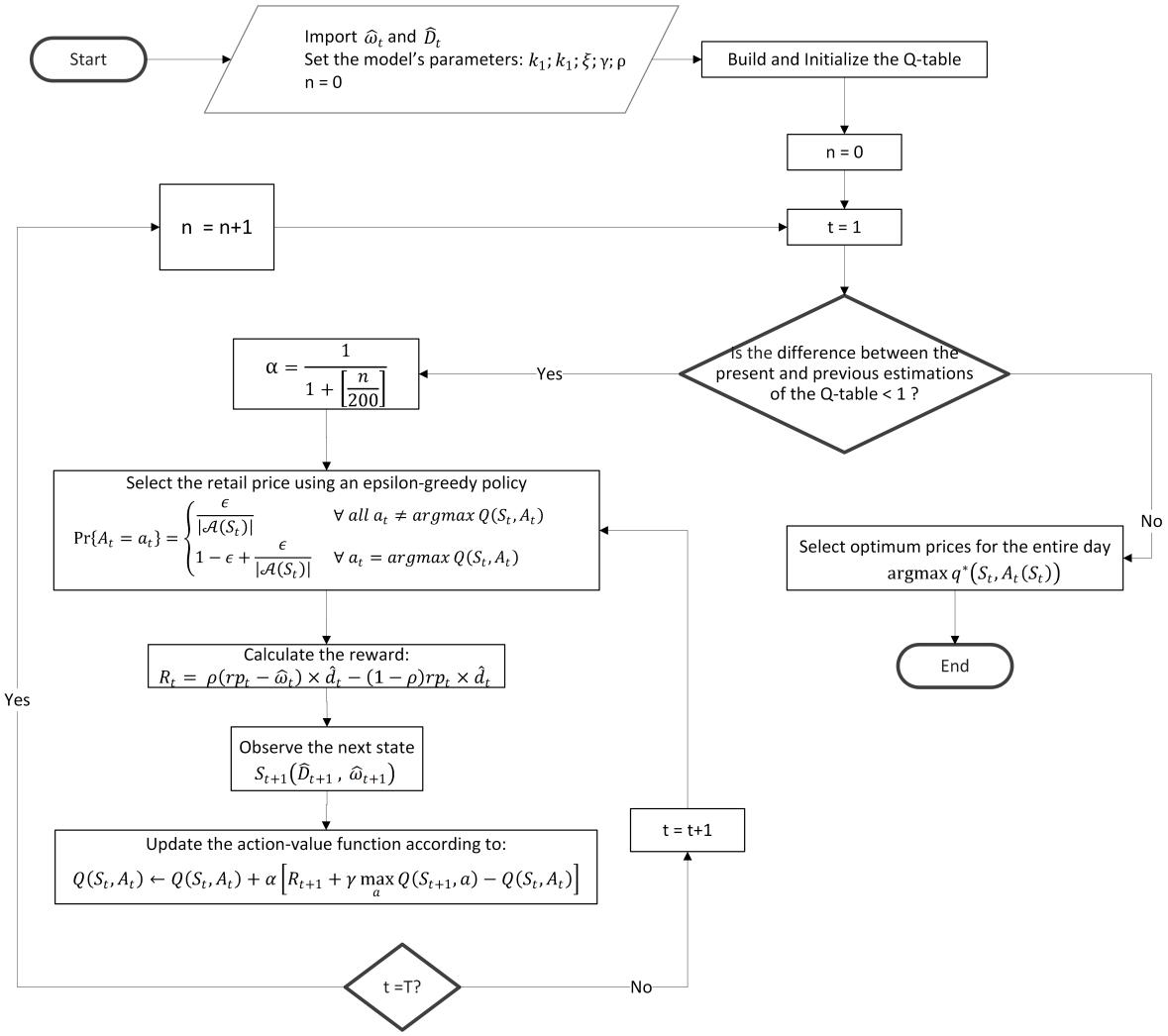
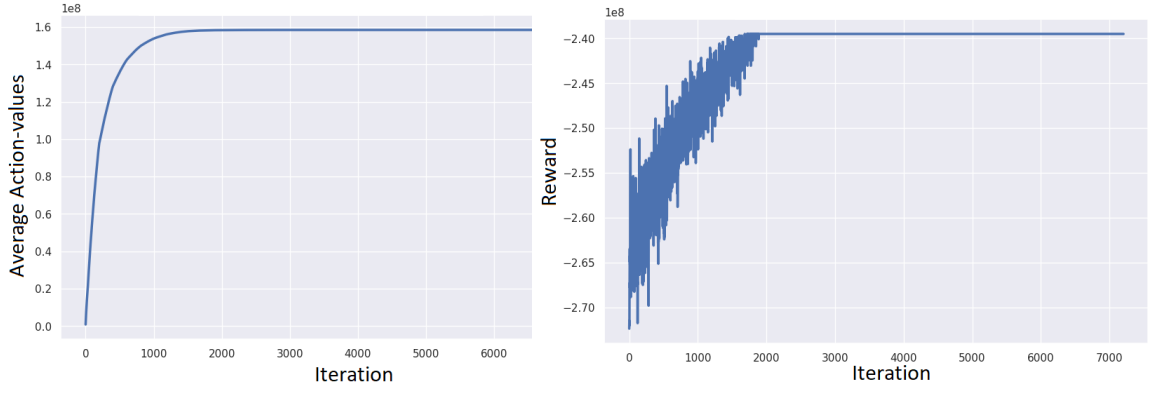


Figure 13 **The proposed Q-learning algorithm**



(a) Average action-values (b) Reward

Figure 14 **The convergence on the 100$^{\text{th}}$ day**

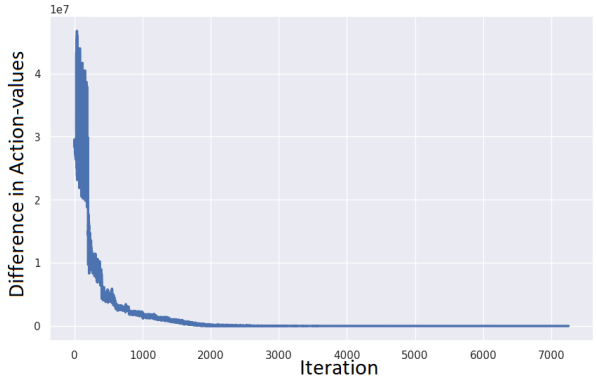


Figure 15 **The difference between present and previous estimations of action-values**

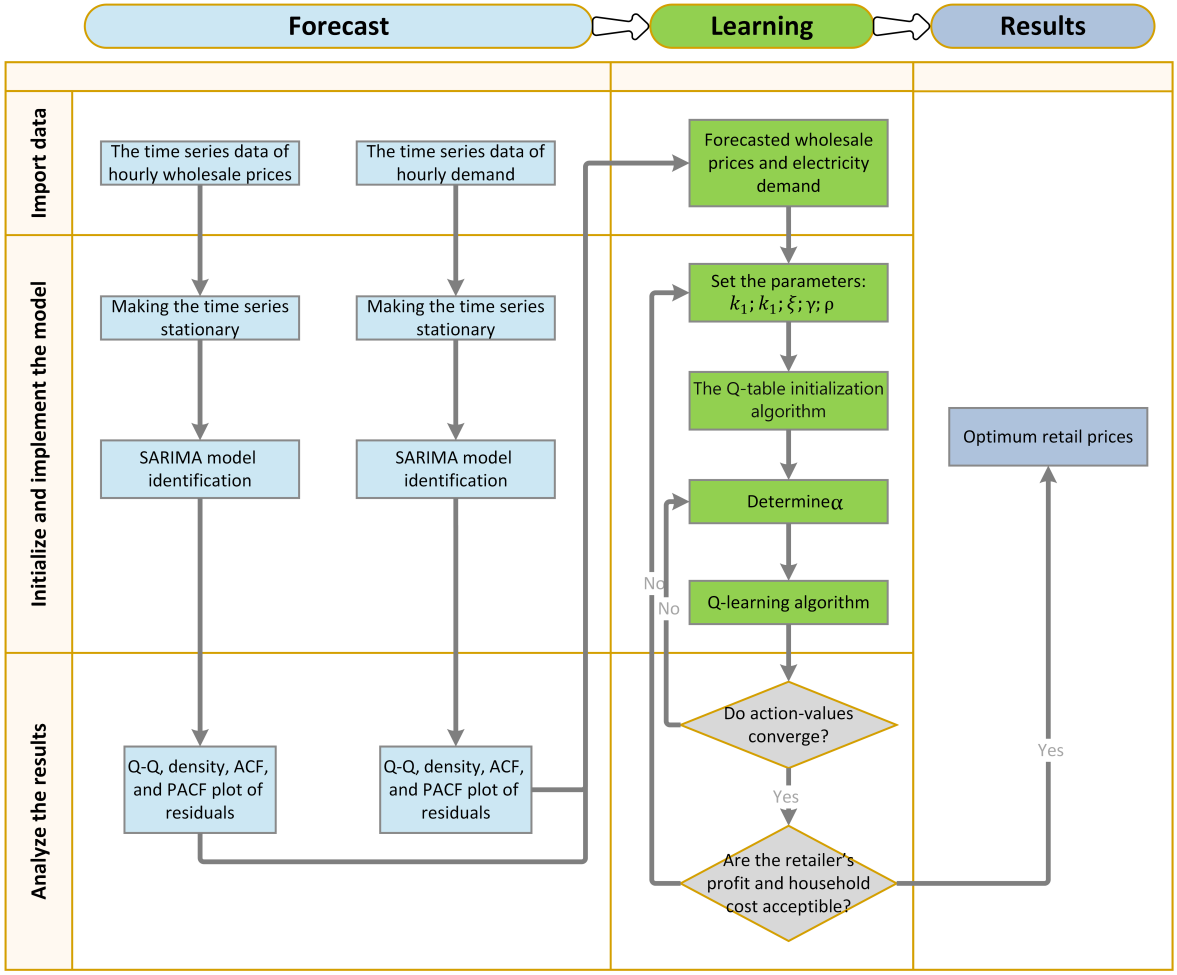


Figure 16 **The proposed framework for electricity dynamic pricing**

5. Results and Analysis

In this case study, the result of the proposed framework outlined in the previous section is the optimum HERPs for the entire fiscal year 2021-2022. Wholesale prices alongside the optimum

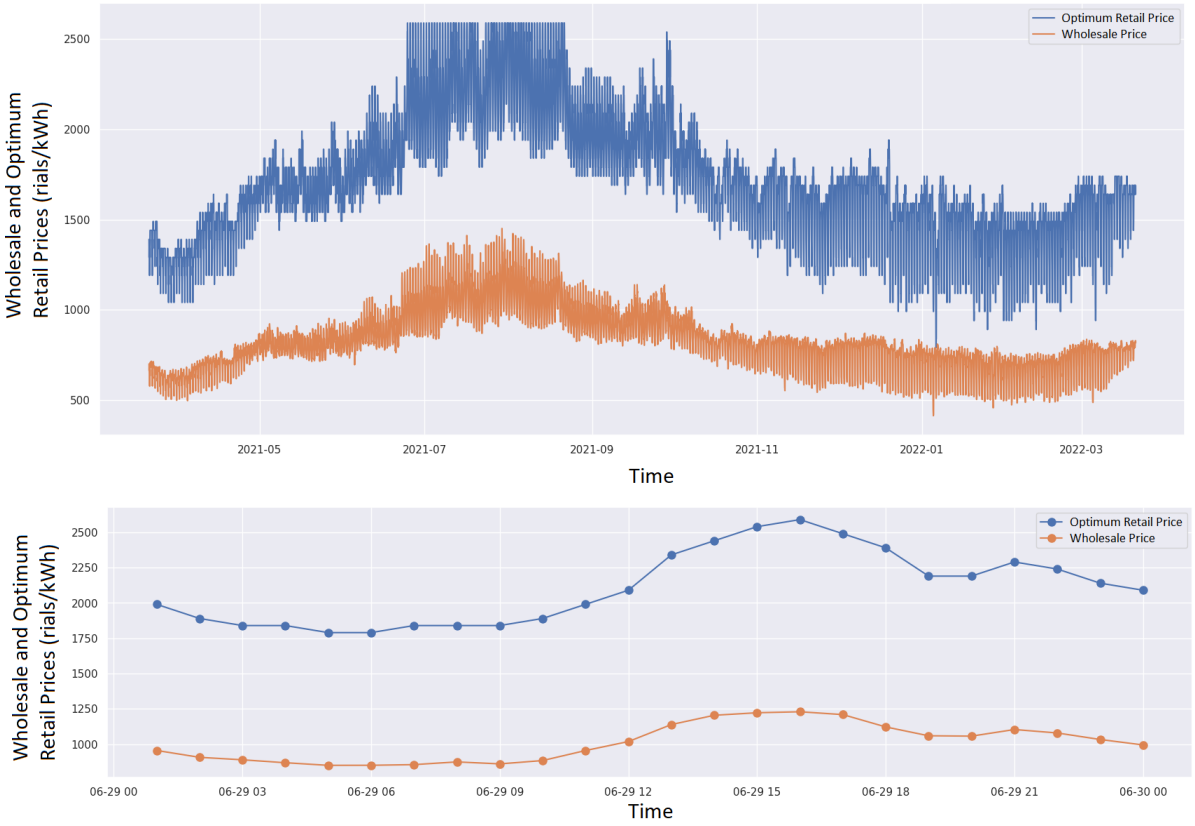


Figure 17 Wholesale prices and the optimum HERPs

HERPs for the entire year and the 100^{th} day of the year are shown in Fig. 17. In compliance with the constraint in Eq. (10), the optimal HERP is expected to be greater than or equal to the corresponding wholesale price at each time step, up to the point where it causes a negative non-critical load.

The demand after and before dynamic pricing implementation for the entire year and the 100^{th} day of the year is shown in Fig. 18, illustrating how dynamic pricing can alter consumption behavior, potentially resulting in reduced consumption, or DR. From a broader perspective, the importance of DR is twofold: peak days and the entire year. Firstly, peak days are of paramount importance due to their very few occurrences throughout the year, whereas the infrastructure needed to meet the demand on those days is extensive and costly. Additionally, this infrastructure often remains idle on all other days. Therefore, the reduction of electricity consumption on peak days can extremely lower the costs associated with electricity supply. Secondly, less power consumption over the entire year means less nonrenewable power generation, leading to less air pollution and less natural resource depletion. Given the prevalence of air pollution in urban centers and its detrimental impact on global warming, decreased power consumption mitigates long-term environmental detriments in this regard. Table 5 provides a comparison of total electricity consumption for the entire year,

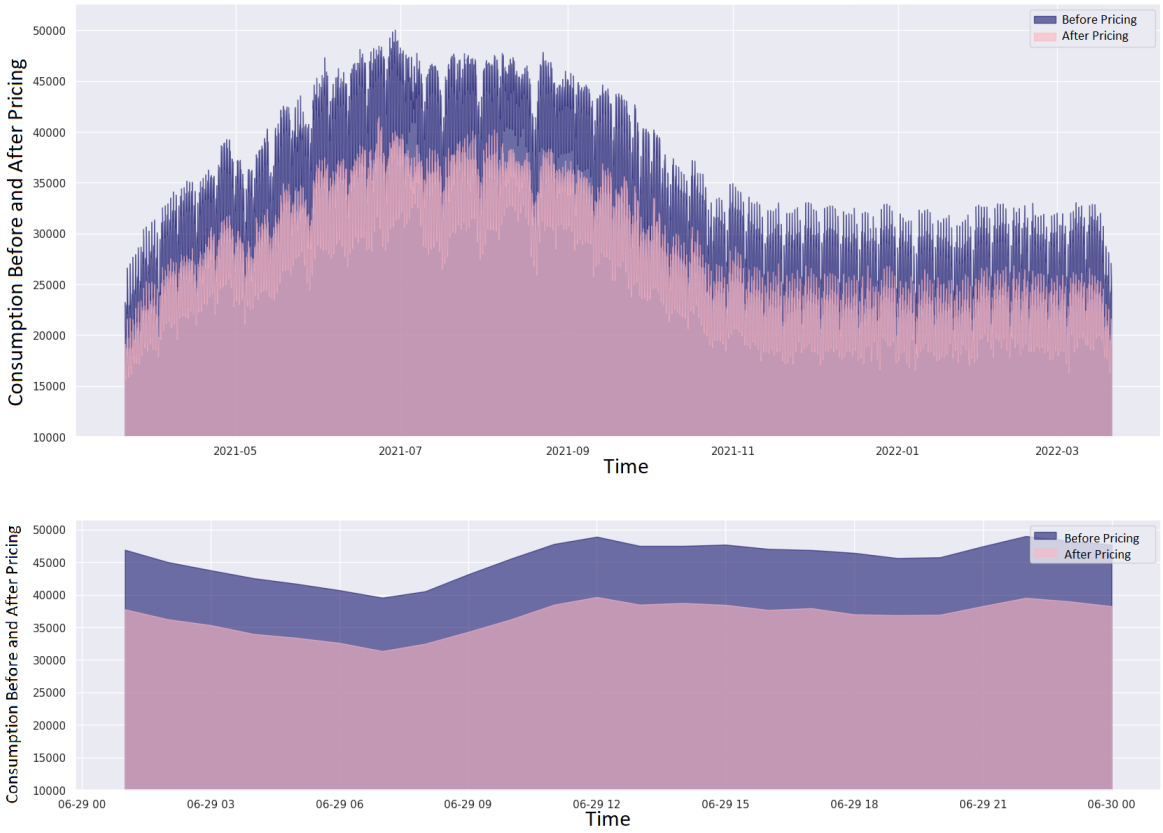


Figure 18 Demand after and before dynamic pricing implementation

Table 5 Total electricity consumption, peak load, and the average HERP analysis

	Total electricity consumption (MWh)	**Peak load (MWh)**	**Average HERP** ($^{\text{IRR}}/_{\text{kWh}}$)
Post-implementation	237495131.96	41486.72	1714
Pre-implementation	294747832.70	50050.30	803
Percentage change	-19.4%	-17.1%	114%

peak load, and the average HERP before and after dynamic pricing implementation. Increasing HERPs by 114% on average leads to an approximately 19% and 17% reduction in total electricity consumption and the peak load, respectively.

Consumption reduction inherently lowers ECs' costs, whereas the increase in HERPs has the opposite effect. Therefore, there are two opposing forces competing with each other to shape ECs' costs. To further investigate this, we proceed to approximate the retailer's profit as well as ECs' costs. The currently implemented retailer's prices are not available in aggregated form because Iran's electricity tariffs are segregated based on geographical regions [32]. According to the "55 Years of Electricity Industry in Iran: A Statistical Reflection" report [44], the average electricity

Table 6 The percentage change in energy CPI

Year	2011	2012	2013	2014	2015	2016
% change in energy CPI	194%	0.8%	1%	21.4%	4.7%	1.5%
Year	2017	2018	2019	2020	2021	
% change in energy CPI	3.4%	3.4%	3.7%	3.4%	3.6%	

Table 7 Estimated Retailer's Profit and ECs' Costs Analysis

	ECs' costs (1000 IRR)	**Retailer's profit (1000 IRR)**
Post-implementation	218689018500	422152708308
Pre-implementation	-15746781394	236644893954

retail price in 2021, indexed to the prices of 2011, is 175 $^{\text{IRR}}/_{\text{kWh}}$. Taking into account the percentage change in consumer price index (CPI) in the energy group obtained from the Central Bank of Iran [45] and listed in Table 6, the present value of the average electricity retail price is calculated to be 803 $^{\text{IRR}}/_{\text{kWh}}$. In light of this value, the retailer's profit as well as ECs' costs are calculated and summarized in Table 7.

According to this calculation, prior to the implementation of dynamic pricing, the retailer was operating at a loss. The main reason for this is the strict regulations on electricity prices imposed by the legislative body in Iran. While these regulations aim to protect ECs, they deprive the retailer of the freedom to set HERPs such that both parties, the retailer and ECs, benefit. Disregarding this restriction and assuming enabling technologies, the retailer could potentially achieve an annual benefit of approximately 219 trillion IRR through dynamic pricing, compared to the current annual loss of approximately 15 trillion IRR. This benefit could be reinvested to build a more comprehensive infrastructure.

In this case study, total ECs' costs increased, meaning that the effect of the rise in HERPs outweighed the effect of consumption reduction. The root cause of this phenomenon can be found in the fact that HERPs were previously so low that the retailer was operating at a loss. Consequently, a dramatic increase in HERPs was necessary for a financial turnaround, making the impact of the rise in HERPs dominant.

One of the key parameters in the model is ρ, as its adjustment can significantly alter the results. We expect that lower values of ρ lead to lower ECs' costs, retailer's profit, and average HERP, as it reflects more importance on ECs' costs. Conversely, its higher values are expected to lead to higher ECs' costs, retailer's profit, and average HERP, since it reflects more importance on the retailer's profit. To investigate its impact on the results, a sensitivity analysis of the parameter ρ is conducted on the 100^{th} day of the year. As shown in Fig. 19, as ρ approaches 1, ECs' costs, the retailer's profit, and the average HERP increase, and vice versa, consistent with our expectations.

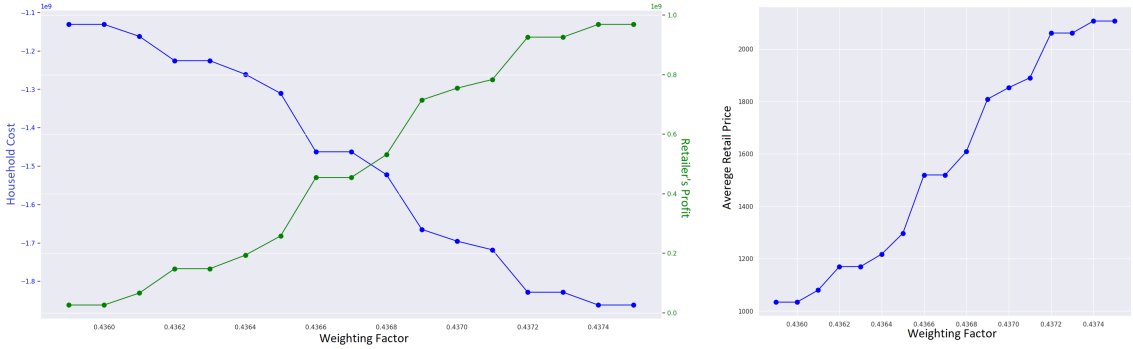


Figure 19 The ECs' costs, retailer's profit and the average HERP vs ρ

6. Conclusions

Population growth, technological advancements, industrial development, and other significant factors have contributed to increased dependence on electricity. However, constructing new power plants and infrastructure is costly, making dynamic pricing one of the best candidates for DR. Additionally, dynamic pricing can benefit both the retailer's profit and ECs' costs. In this study, we proposed a framework for determining the optimum HERPs to maximize the retailer's profit while minimizing ECs' costs simultaneously. To investigate its effectiveness, this framework was implemented on Iran's power grid in the fiscal year 2021-2022 as a case study. Firstly, wholesale prices were forecasted for each day, and then a reinforcement learning approach was used to find the optimum pricing policy.

Day-ahead forecasting for wholesale prices was conducted using the SARIMA method, one of the most widely used algorithms for time-series analysis. This method allows for the consideration of non-stationarity, which is evidently present in wholesale prices. Forecasted demand, provided by the Iran Grid Management Company, and forecasted wholesale prices were used as inputs for Q-learning. The system was modeled as an MDP, and optimum HERPs were obtained using Q-learning. The convergence of the algorithm was then investigated. The results indicate that by increasing HERPs by approximately 114%, average consumption and peak load decreased by 19.4% and 17.1%, respectively, considering a 20% non-critical load and $\rho = 0.5$. Additionally, the retailer achieved an annual benefit of approximately 219 trillion IRR through dynamic pricing, compared to the current annual loss of approximately 15 trillion IRR. Moreover, the model was found to be significantly sensitive to ρ, and as it approaches 1, ECs' costs, the retailer's profit, and the average HERP increased, and vice versa.

The managerial insight derived from this study is as follows: Using the framework proposed in this study, the market regulator can start dynamic pricing for electricity by assigning more weight to ECs, leading to lower HERPs, and gradually move to its preferred weighting factor.

The limitations of this study include the absence of enabling technologies in Iran, which hinders the implementation of dynamic pricing. Additionally, retailers cannot decide on their pricing policy independently, as they have to abide by the legislative body. Moreover, an accurate estimation of the price elasticity of demand is not available. For future research, using more advanced time-series methods, such as autoregressive conditional heteroscedasticity (ARCH) or generalized autoregressive conditional heteroscedasticity (GARCH), is recommended. These methods are ideal for analyzing volatile data, like wholesale prices. Furthermore, if geographically segregated data were available, a more accurate analysis could be conducted and more valuable insights could be derived.

References

[1] A. Miglani, N. Kumar, V. Chamola, and S. Zeadally, "Blockchain for internet of energy management: Review, solutions, and challenges," *Computer Communications*, vol. 151, pp. 395–418, 2020.

[2] D. Stanelyte, N. Radziukyniene, and V. Radziukynas, "Overview of demand-response services: A review," *Energies*, vol. 15, no. 5, p. 1659, 2022.

[3] A. V. Den Boer, "Dynamic pricing and learning: historical origins, current research, and new directions," *Surveys in operations research and management science*, vol. 20, no. 1, pp. 1–18, 2015.

[4] R. Phillips, *Pricing and Revenue Optimization*. Stanford University Press, 2005. [Online]. Available: https://books.google.com/books?id=InuQPrC6GtQC

[5] G. Dutta and K. Mitra, "A literature review on dynamic pricing of electricity," *Journal of the Operational Research Society*, vol. 68, no. 10, pp. 1131–1145, 2017.

[6] A. Faruqui and J. Palmer, "The discovery of price responsiveness–a survey of experiments involving dynamic pricing of electricity," *Available at SSRN 2020587*, 2012.

[7] P. Warren, "A review of demand-side management policy in the uk," *Renewable and Sustainable Energy Reviews*, vol. 29, pp. 941–951, 2014.

[8] J. S. Vardakas, N. Zorba, and C. V. Verikoukis, "A survey on demand response programs in smart grids Pricing methods and optimization algorithms," *IEEE Communications & Surveys Tutorials*, vol. 17, no. 1, pp. 152–178, 2014.

[9] Q. Qdr, "Benefits of demand response in electricity markets and recommendations for achieving them," *US Dept. Energy, Washington, DC, USA, Tech. Rep*, vol. 2006, 2006.

[10] K. Kostková, L. Omelina, P. Kyčina, and P. Jamrich, "An introduction to load management," *Electric Power Systems Research*, vol. 95, pp. 184–191, 2013.

[11] C. Yang, C. Meng, and K. Zhou, "Residential electricity pricing in china: The context of price-based demand response," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 2870–2878, 2018.

[12] P. L. Joskow and C. D. Wolfram, "Dynamic pricing of electricity," *American Economic Review*, vol. 102, no. 3, pp. 381–85, 2012.

[13] J. D. Quillinan, "Pricing for retail electricity," *Journal of Revenue and Pricing Management*, vol. 10, no. 6, pp. 545–555, 2011.

[14] C. Joe-Wong, S. Sen, S. Ha, and M. Chiang, "Optimized day-ahead pricing for smart grids with device-specific scheduling flexibility," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 6, pp. 1075–1085, 2012.

[15] T.-C. Chiu, C.-W. Pai, Y.-Y. Shih, and A.-C. Pang, "Optimal day-ahead pricing with renewable energy for smart grid," in *2014 IEEE International Conference on Communications Workshops (ICC)*. IEEE, Conference Proceedings, pp. 472–476.

[16] G. Pecoraro, S. Favuzza, M. Ippolito, G. Galioto, E. R. Sanseverino, E. Telaretti, and G. Zizzo, "Optimal pricing strategies in real-time electricity pricing environments: An italian case study," in *2015 International Conference on Clean Electrical Power (ICCEP)*. IEEE, Conference Proceedings, pp. 376–381.

[17] J. Jiang, Y. Kou, Z. Bie, and G. Li, "Optimal real-time pricing of electricity based on demand response," *Energy Procedia*, vol. 159, pp. 304–308, 2019.

[18] S. Joseph and E. Jasmin, "Demand response program for smart grid through real time pricing and home energy management system," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 5, pp. 4558–4567, 2021.

[19] B. Xu, J. Wang, M. Guo, J. Lu, G. Li, and L. Han, "A hybrid demand response mechanism based on real-time incentive and real-time pricing," *Energy*, vol. 231, p. 120940, 2021.

[20] I. Antonopoulos, V. Robu, B. Couraud, D. Kirli, S. Norbu, A. Kiprakis, D. Flynn, S. Elizondo-Gonzalez, and S. Wattam, "Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review," *Renewable and Sustainable Energy Reviews*, vol. 130, p. 109899, 2020.

[21] H. Xu, H. Sun, D. Nikovski, S. Kitamura, and K. Mori, "Learning dynamical demand response model in real-time pricing program," in *2019 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, Conference Proceedings, pp. 1–5.

[22] T. Lu, X. Chen, M. B. McElroy, C. P. Nielsen, Q. Wu, and Q. Ai, "A reinforcement learning-based decision system for electricity pricing plan selection by smart grid end users," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2176–2187, 2020.

[23] B.-G. Kim, Y. Zhang, M. Van Der Schaar, and J.-W. Lee, "Dynamic pricing for smart grid with reinforcement learning," in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, Conference Proceedings, pp. 640–645.

[24] ——, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Transactions on smart grid*, vol. 7, no. 5, pp. 2187–2198, 2015.

[25] R. Lu, S. H. Hong, X. Zhang, X. Ye, and W. S. Song, "A perspective on reinforcement learning in price-based demand response for smart grid," in *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, Conference Proceedings, pp. 1822–1823.

[26] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach," *Applied Energy*, vol. 220, pp. 220–230, 2018.

[27] S. Z. M. Ziabari, "Demand prediction for dynamic pricing in smart electricity markets," Thesis, 2020.

[28] Y. Wan, J. Qin, X. Yu, T. Yang, and Y. Kang, "Price-based residential demand response management in smart grids: A reinforcement learning-based approach," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 1, pp. 123–134, 2021.

[29] R. Bagherpour, N. Mozayani, and B. Badnava, "Optimizing dynamic pricing demand response algorithm using reinforcement learning in smart grid," in *2020 25th International Computer Conference, Computer Society of Iran (CSICC)*. IEEE, Conference Proceedings, pp. 1–5.

[30] L. Zhang, Y. Gao, H. Zhu, and L. Tao, "A distributed real-time pricing strategy based on reinforcement learning approach for smart grid," *Expert Systems with Applications*, vol. 191, p. 116285, 2022.

[31] (2023) Iran Grid Management Company. [Online]. Available: www.igmc.ir

[32] (2021) Electricity tarrifs and their general conditions. Ministy of Energy. [Online]. Available: https://tariff.moe.gov.ir

[33] (2022) Descriptive statistics of iran's electric power industry from 2011-2021. Iran organization for Management of Electric Power Generation and Transmission. [Online]. Available: https://amar.tavanir.org.ir

[34] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.

[35] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[36] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[37] A. Amini Fard and S. Estedlal, "Estimation of the residential electricity demand function in iran."

[38] H. Aalami, M. P. Moghaddam, and G. Yousefi, "Demand response modeling considering interruptible/curtailable loads and capacity market programs," *Applied energy*, vol. 87, no. 1, pp. 243–250, 2010.

[39] E. Nia, Mohammadi, Teymoori, Zamani, and Aboutaleb, "The evolution of price elasticity of residential electricity demand in iran using a kalman filter approach," *Financial Economics (Iran)*, vol. 25, no. 7, pp. 147–176, 2013.

[40] S. A. Jalaee, S. Jafari, and S. Ansari Lari, "The estimation of electricity consumption in the residential sector in iran: A provinces panel," *Iranian Energy Economics*, vol. 2, no. 8, pp. 69–92, 2013. [Online]. Available: https://jiee.atu.ac.ir/article_700_720bbac1e4b447548aca965131d421ed.pdf

[41] N. Abbaszadeh, M. Qavami, and A. Bahmani, "Price elasticity of electricity demand in ıran based on computable general equilibrium model," *Journal of Accounting and Marketing*, 2014.

[42] a. nazemi, s. ghaderi, and s. tojar, "The effect of price elasticity on demand response program modeling in the smart power grid," *Quarterly Journal of Industrial Economics Research*, vol. 5, no. 17, pp. 29–44, 2021. [Online]. Available: https://indeco.journals.pnu.ac.ir/article_8354.html

[43] M. Vesal, A. Cheraghi, and M. Rahmati, "Household electricity demand estimation: a regression discontinuity design approach," *Quarterly Journal of Economic Research and Policies*, vol. 29, no. 99, pp. 7–57, 2021. [Online]. Available: http://qjerp.ir/article-1-3032-fa.html

[44] (2022) 55 years of electricity industry in iran: A statistical reflection. Iran organization for Management of Electric Power Generation and Transmission. [Online]. Available: https://amar.tavanir.org.ir/pages/omomi/55saleomomi.pdf

[45] (2023) The consumer price index for goods and services in urban areas of iran. Central Bank of Iran. [Online]. Available: https://www.cbi.ir/