Siep Società italiana di economia pubblica

XVI CONFERENZA

POLITICHE PUBBLICHE, SVILUPPO E CRESCITA

Pavia, Università, 7 - 8 ottobre 2004

PEAK-LOAD PRICING IN THE ELECTRICITY MARKET: THEORY AND PRACTICE

GRAZIANO ABRATE

pubblicazione internet realizzata con contributo della COMPAGNIA di San Paolo

società italiana di economia pubblica

dipartimento di economia pubblica e territoriale – università di Pavia

PEAK-LOAD PRICING IN THE ELECTRICITY MARKET: THEORY AND PRACTICE[♣]

Graziano Abrate
(*University of Pavia*, *Ceris-CNR*, *HERMES*)
Via Avogadro 8
10121 Torino
Tel: +39-011-5601210
Fax: +39-011-5626058
Email: g.abrate@ceris.cnr.it

Abstract

A peak-load problem arises when a commodity is characterised by non-storability and periodical (daily, weekly, seasonal) demand fluctuations. This situation fits to the electricity market, as well as to many other public utility goods and services (e.g. phone calls, transport). Economists have traditionally indicate price differentiation over time as the theoretical optimal solution, since the pioneer work of Steiner (1957), which provided the basis for the development of a wide strand of literature going under the name of peak-load pricing theory. The approach evolved to consider stochastic demand and supply, and towards the idea of pricing electricity in real time. The idea has somehow been implemented in the deregulated electricity market, with the introduction of the power exchange spot market. This paper aims at describing the theoretical foundations for the existence of a spot market, comparing the critical theoretical assumptions with the practical implementation. The classical peak-load literature will be linked to a new strand of literature which has focused on retail electricity competition. Finally, the theoretical solutions on optimal pricing will be compared with the practice.

[♣] I am grateful to Alberto Cavaliere for useful comments on an earlier version of the work.

1. Introduction

Demand for electricity presents sharp periodical variations, whereas production is subject to rigid short term capacity constraint. During off-peak times, there is plenty of capacity and the cost of producing an additional kilowatt-hour only reflects fuel and some operating and maintenance costs. On the other hand, during peak periods, the capacity constraint will be binding and the incremental cost can be significantly higher. Economists have traditionally shown great interest on the peak load problem, that electricity shares with other non-storable goods, and whose theoretical solution calls on price differentiation over time, sometimes together with some form of rationing.

In the electricity market, the topic is particularly relevant since many countries have undergone structural reforms moving from a model in which there was a vertically integrated unit with monopoly power towards a deregulated market. In such a restructured model, the idea of spot pricing has been implemented. Therefore, in the wholesale market, prices vary on a hourly (or half-hourly) basis, reflecting the interaction between demand and supply. However, the end-use consumer generally faces a fixed retail price, thus independent from the wholesale price and the actual system load: consequently, demand is almost completely inelastic and does not play an active role in determining wholesale prices. Lack of demand participation can favour market power behaviour of generators and price volatility: the classical example is represented by the experience in California during the summer of 2000. A new strand of literature has been focused on the ways to promote demand responsiveness.

The aim of this chapter is reviewing the main findings in the literature tracing its evolution and trying to link theory with practical issues. In particular, the objective is to evaluate which can be the role of time-varying tariffs in enhancing the demand participation in the electricity spot wholesale market. After giving some basic definitions concerning the demand in the power market (section 2), I will describe the theory of peak load pricing, starting from the pioneer work of Steiner (1957) and describing the developments to the basic model (section 3). I will show how the theory evolved towards the definition of real time pricing, explaining its role as instruments of demand management (Section 4). In section 5, I will briefly describe the feature of a restructured electricity market and I will link the basic peak load theory, concerning regulated public utilities, with the new strand of literature that has developed over the question of retail competition. Particular attention will be given to the consequences of having a part of customers which is price-insensitive. The pure market clearing approach will be compared with solutions allowing for rationing. Section 6 moves on to the description of time-varying schemes that have been implemented in practice. I will explain what is meant by demand response programs and the different ways they can be implemented to enhance demand responsiveness, highlighting their possible effect on welfare and their effectiveness in the sense of actually achieving demand responsiveness. I will focus in particular on the differences between adopting Real Time Pricing (RTP) and Time of Use Pricing (TOU). The key difference is that under RTP prices adjust frequently according to the actual balance between demand and supply, while TOU provide *preset* tariffs, and so they are less likely to reflect the prices in the wholesale market. In Section 7 I give some concluding remarks and describe some directions for future research.

2. Demand-side economics in electricity markets: some basic definitions

The physical aspects of supply and demand must receive a great attention for understanding the fundamental economics of markets. For power, Stoft (2002) underlines the peculiar role played by the shifts in the level of demand that are *not* associated with price. Indeed, demand is highly variable between and within a day, and these hourly fluctuations determine the key long-run characteristics of supply. Traditionally, the demand for power can be described by a *load-duration curve*, which measures the number of hours per year the total load is at or above any given level of demand. Even if this curve does not include information on the sequence of the load levels[1], it gives information about the peak-level demand and its duration (say, the peak demand was 1,211 MW; the demand was above 1,100 MW for 122 hours in the year; and so on). A natural interpretation for such data is the probability that load will be at or above a certain level (in the previous example, 122 out of 8,760 hours in a year, i.e. 1.4 per cent of probability that demand will exceed 1,100 MW).These data are very important in designing the productive structure, because since electricity is *not storable*, supply is equal to consumption at any time (ignoring losses)[2]. Therefore, peak demand must be satisfied by production from generators that are used as little as 1% of the time. The technology used to build such generators, so-called *peakers*, significantly different from the one used for the *baseload* generators, which run most of the time, and, in particular, the first ones generally imply a higher marginal cost of production[3]. With a very broad approximation, it could be said that a higher load level is associated to a

[1] So, for example, the same curve can describe wide daily swings in demand and little seasonal variation or wide seasonal variation and limited daily swings.

[2] To be precise, the amount stored is minuscule and cannot be utilized for trade.

[3] Actually startup costs of plants should be considered. Plants with high startup cost may be running in spite of other plants which are cheaper strictly from the point of view of marginal operating cost.

higher marginal cost, which can greatly increase when demand is at the highest level. It must also be noted that, even if supply always equals *consumption*, it may not equal *demand*: since supply is subject to rigid short term capacity constraints, demand may be higher than the maximum possible supply in a certain moment[4].

Another way to characterise the time-frame allocation of consumption is the *load profile*, which refers to the percentage of consumption allocated to a certain time-period with respect to the total consumption (e.g. the percent distribution of consumption between day and night). While the load-duration curve is mainly useful to define the aggregate needs and the choices in terms of investments, the load profile provides a more accurate measure of the time-frame preferences and is more relevant when talking about individual behaviour. In fact, consumers exhibit different load profiles; for example, the individual consumption peak does not necessarily correspond to the system peak.

All above definitions are independent from any consideration about *prices*, a dimension that must be added when talking about a market. Price responsive behaviour lies at the basis of economic theory, and here it is useful to recall some basic concepts of demand-side economics applied to this specific context. Suppose that there are only two time-period, namely a peak period and a non-peak one. In this case, we need to specify two distinct demand function, one for the peak period (x_p) and one for the off-peak (x_{op}), treating the electricity consumed in the peak hours and the electricity consumed off-peak as they were two different commodities. In this context, each demand will depend on both prices, p_p and p_{op} (and other variables), and it is necessary to define two different types of price elasticity:

a) the *own-price* elasticity relates the variation in the demand to a change in its own price (for example, the variation in the peak consumption induced by a change in the peak price).

b) the *cross-price* elasticity gives the sensitivity of the demand to the price charged in the other time-period (for example, the sensitivity of the peak demand to a change in the off-peak price). In this case, relative prices matter.

It is important to explain the different meaning of these two types of price responsive behaviour a consumer may exhibit in the electricity market. The own-price elasticity represents the consumers willingness to curtail or increase consumption as a

[4] Technically, the difference between supply and demand cannot be indicated by flows of power, but must be measured in terms of voltage and frequency. Demand for power is defined as the amount of power that would be consumed if system frequency and voltage were equal to their target values for all consumers. If voltage or frequency are low, then customers consume less power than they would like so supply is less than demand. For a more detailed explanation see Stoft (2002), pag. 40-48 and pag. 373-388.

function of higher or lower prices; instead, the cross-price elasticity corresponds to the willingness to shift load from peak hours to off-peak hours in response to price, while keeping overall consumption the same. The latter is therefore related with a modification of the consumers load profile. Of course, consumers may be heterogeneous in terms of both load profiles, own- and cross-price elasticities. Moreover, these values may vary on a daily, weekly and seasonal basis.

3. The theory of peak-load pricing

A *peak-load* problem arises when a commodity is characterised by *non-storability* and periodical (daily, weekly, seasonal) *demand fluctuations*. In the previous section I described how this situation fits to the electricity market; indeed, it can refer also to many other public utility goods and services (e.g. phone calls, transport). The common problem in such industries lies in the need of installing a capacity large enough to meet demand at the peak, indeed under-utilised during the remainder of the cycle. Price differentiation over time has traditionally been indicated as a valid instrument to mitigate this inefficiency. In this section I describe the classical *peak-load pricing* solution (Boiteaux, 1949; Steiner ,1957) and the following extensions.

3.1. The origin of peak-load pricing theory

The peak-load problem originates in the context of regulated industries with reference to the need of covering the capacity costs with an appropriate tariff design. The early debate[5] focused on defining cost-based pricing mechanisms, and in particular a solution based on long-run marginal cost pricing is proposed by Houthakker (1951). A more precise definition of the peak-load problem can be found in Steiner (1957): "…to find an appropriate price policy that leads to the *correct amount of physical capacity* and its efficient utilisation, and that also *covers the full social costs* of the resources used". Steiner (1957), generally recognised as the originator of the peak-load pricing theory, showed that purely cost-based prices are not efficient, and that the theoretical solution requires the explicit consideration of the demand behaviour. In his model, the following framework is considered:

(a) costs are linear and only 1 technology is available: b is the operating marginal cost and β is the per-day cost of providing a unit of capacity.

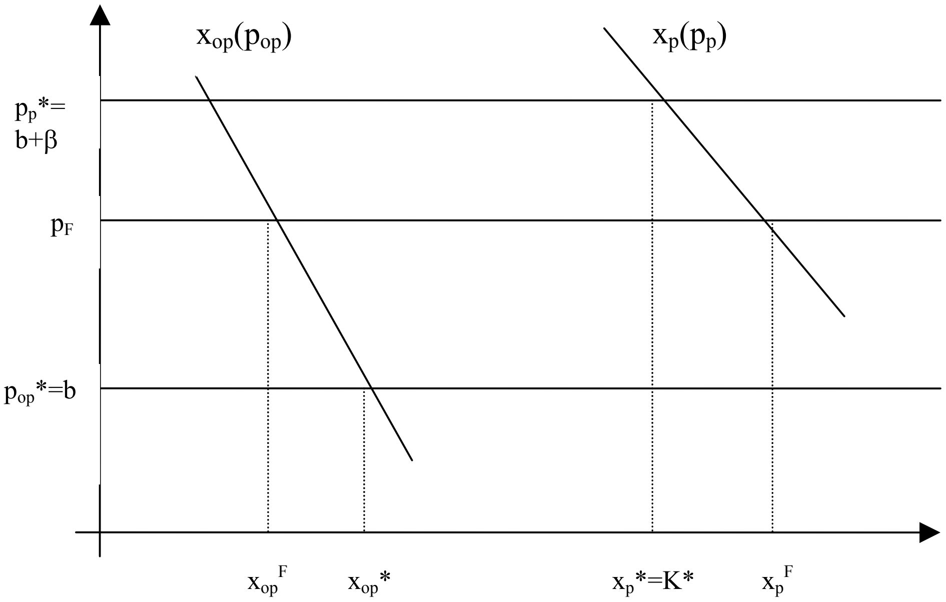
[5] A first treatment of the peak-load problem goes back at least to the work of Bye (1926).

(b) there is no uncertainty on demand and supply (i.e. the peak-load is *deterministic*);

(c) demand in each period is given as a continuous and declining function of price;

Figure 1 illustrates the simplified case of 2 time-periods (a peak demand, x_p, and an off-peak demand, x_{op}, which in this case depend only on their own price, p_p and p_{op})[6]. Under the above assumptions, and as long as $x_{op}(\mathrm{p_{op}}^*) < x_p(\mathrm{p_p}^*)$, off-peak consumers only pay for operating costs, while peak consumers determine the efficient level of capacity and pay for it. This is nothing but the solution of long-run marginal cost pricing; however, Steiner recognised the possibility that pricing at marginal cost can shift the peak and off-peak periods (i.e. $x_{op}(\mathrm{b}) > x_p(\mathrm{b}+\beta)$). In this case, the optimal solution requires equal outputs during among peak and potential peak period (in the case of only 2 time-period, this also implies fully-utilised capacity).

Figure 1. A classical peak-load problem



3.2. Extensions of the basic model

The theory has progressively investigated on the effects of relaxing the above assumptions (a), (b), (c), to introduce a more complex and realistic framework, without seriously undermining the insights from the basic model.

[6] Considering a higher number of time-periods does not vary the main insights of the theory.

Table 1. Optimal prices and capacity under different assumptions.

Author	Main assumptions of the model	Results for price and capacity
Steiner (1957)	No uncertainty on demand and supply (deterministic peak load) Demands are taken as given functions of prices. Costs are linear and there is only 1 technology available: *b* is the operating marginal cost and β is the per-day cost of providing a unit of capacity	1) $p_{op}^* = b$ $p_p^* = b + \beta$ $K^* = x_p(b + \beta)$ as long as $x_{op}(p_{op}^*) < x_p(p_p^*)$ 2) If this condition does not hold, then the shifting peak case arise: $p_{op}^* + p_p^* = 2b + \beta$ $p_p^* > p_{op}^*$ $K^* = x_p(p_p^*) = x_{op}(p_{op}^*)$
Crew and Kleindorfer (1976)	Extension to multiple technologies. N technologies are available, such that: $\beta_1 > \beta_2 > \ldots > \beta_N$ $b_1 < b_2 \ldots < b_N$	For the two technology case: $b_1 < p_{op}^* < b_2 < b_2 + \beta_2 = p_p^* < b_1 + \beta_2$ More options in technology imply therefore a *lower peak price* and a *higher off-peak price*
Kleindorfer and Fernando (1993)[7]	Extension to take into account uncertainty on demand and supply. This involves possibility of outage and the need for rationing. Endogenous determination of the optimal level of reliability.	Results similar to the deterministic case, but with short-run marginal cost including the expected outage cost. 1 period, 1 technology example: $p^* = b + \beta/a + \Lambda$ where: *a* = availability factor Λ = excess of willingness to pay over price for unserved energy to the marginal consumer
Shy (2001)	Demand in different periods are not independent. Endogenous choice of consuming during the peak or during the off-peak period. Introduction of a time discount factor (ρ) and a "flexibility" index (δ)	$p_{op}^* = b$ $p_p^* = b + \beta(1+\rho)/2$ $K^* = x_p(p_p^*)$ Peak price is lower (and optimal capacity greater) than in the basic case (unless ρ=1)

[7] Built on the basis of previous works: Brown and Johnson (1969), Visshser (1973)

(a) A first area of extensions has been dealt with a more accurate description of costs and technology. Williamson (1966) expanded the framework to include indivisible capacity increments. Crew and Kleindorfer (1976) considered the implications of having more than one type of technology to meet demand. This is typical in electricity markets, where baseload and peakers plants have different characteristics both in terms of investment and operating costs ($\beta_{\text{base}} > \beta_{\text{peak}}$; $b_{\text{base}} < b_{\text{peak}}$). The results show that having diverse technology leads to a reduction in the optimal peak price and an increase in the off-peak price.

(b) When it comes to *stochastic* realisations of demand and supply (Brown and Johnson, 1969; Visshser, 1973; Kleindorfer and Fernando, 1993), a *reliability* problem arises, a common issue when talking about electricity. Since it may be the case that demand exceeds the supply, some forms of *rationing* may be required at the social optimum. It is important to recall that for the electricity services an unforeseen state of excess demand means blackout, with consumption equal to 0 for *all* consumers. Therefore, the stochastic framework calls on the need of demand management instruments, such as time and space differentiated pricing, interruptible tariffs, and other means of reducing the probability of outage and the welfare losses in excess demand states. These arguments will be treated more in deep in the following sections. Here it is useful to recall the basic effects of introducing randomness in the peak-load model. As shown in Kleindorfer and Fernando (1993), prices in each time-period should be set equal to the *expected* short-run marginal cost, which must include the *expected outage costs*. Optimal capacity minimises expected costs of operation, and equate marginal curtailment cost to the peaker marginal capacity cost.

(c) Last area of improvements refers to the explicit modelling of customer choice, thus starting from the utility function instead of the demand function. Shy (2001) endogenises the choice of consuming in the peak or in the off-peak period, in a dynamic setting. Each consumer *i* (who is assumed to buy only once) decides whether to buy or to postpone consumption, on the basis of relative daily (weekly, seasonal) pricing of the service. The choice also depends on a parameter indexed by δ_i on the interval [0,1], which introduces customer heterogeneity and can be interpreted as the flexibility of consumer in switching period of consumption according to price changes. The model also introduces a time discount factor (ρ), which reduce the utility level associated to postponed consumption. Only this last hypothesis (if $\rho \neq 1$) affect the optimal outcome, resulting in a lower peak price with respect to the basic framework. However, this modification seems not to be relevant when considering daily substitution in electricity markets (ρ=1 should be a reasonable assumption).

Table 1 provides a summary of the results on optimal prices and capacity under different assumptions, with reference to the case of 2 time-periods. All the literature described above answer to the question of optimal pricing from the point of view of a regulator with perfect information on cost structure and demand[8]. As pointed out by Crew et al. (1995), an underlying common approach to derive efficient prices can be defined, which follows from the maximisation of an explicit social welfare function.

$$\underset{p}{\text{Max}}\, W = TR + S - TC \qquad [1]$$

where W is the net social benefit, given by the sum of producer surplus (TR-TC, total revenue – total costs) and consumers' surplus (S). [1] is typically constrained by a breakeven constraint for the production sector. Indeed, peak-load pricing can be viewed as a form of Ramsey pricing: the peculiarity of peak-load analysis is that the welfare maximisation refers to the provision of a vector of products differentiated only by the time of consumption.

A separable form is used to represent the preferences of consumers:

$$U(x,m,\theta) = V(x,\theta) + m\,, \qquad \theta \in \Theta \qquad [2]$$

where *x=(x*$_1$*, ..., x*$_T$*)* is the vector of goods supplied by the regulated sector (i.e. the consumption of electricity in the different time-periods) and *m* is an Hicksian aggregate representing the utility from all other goods. θ is a parameter that allows for consumers' heterogeneity, with $f(\theta)$ being the density of consumers of type θ.

The Ramsey problem can be stated as:

$$\underset{p\geq 0}{\text{Max}}\, W(p) = \int_{\theta}\left[V(x(p,\theta),\theta - \sum_{T} p_i x_i(p,\theta) \right] f(\theta)d\theta + \Pi(p) \qquad [3]$$

subject to $\Pi(p) = \sum_{T} p_i x_i(p) - C(x) \geq \Pi_0$

where *C(x)* is the cost function and Π_0 is some desired profit level (e.g. 0).

The solution of [3] yields the first-best price schedule[9]:

$$\sum_{j\in T} \frac{p_j - c_j}{p_j} \eta_{ij} = -\kappa \qquad \forall i \in T \qquad [4]$$

where η_{ij} is the cross-elasticity between consumption in two different periods, and κ is the so-called Ramsey number, which is positive except when the profit constraint is not binding. [4] implies that, as long as products are substitutes over time (η_{ij}>0, for

[8] In the case of stochastic realisations of supply and demand, the perfect information is referred to the knowledge of probability structure.

[9] In the sense that, when coupled with appropriate lump-sum transfers, the Ramsey solution can Pareto dominate every other linear price schedule and lump-sum transfer schedule satisfying the profit constraint. See Crew et al. (1995) for the analytical solution of [3].

i≠j), price will always exceed marginal cost in all period, except at the unconstrained welfare optimum.

4. Real-time pricing and demand management

In the previous section, we have seen that stochastic realisations of supply and demand may results in states of excess demand. In such a situation, price differentiation becomes an instrument of *demand management*, which can be used to reduce the probability of having unforeseen blackouts. In a deterministic setting, price responsive behaviour affects the optimal price schedule and the determination of the efficient level of industry capacity. However, as far as the efficient cost allocation is concerned, price differentiation could be an optimal solution even if price elasticity was zero. It is clear that when price differentiation becomes also an instrument of demand management, the degree of price responsiveness assumes a greater relevance. Besides, the problem of optimal pricing takes a *dynamic* aspect, since it would be necessary to adjust tariffs instantly, to take into account of the stochastic variations in the demand-supply balance. The concept of pricing public utility services in *real time* was first introduced by Vickrey (1971), who called them "responsive prices" and argued that this yields a first-best outcome in a world where there are no transaction costs, customers are risk neutral and can respond optimally to price signals. In fact, real time pricing implies the solution (at least partial) of the uncertainty concerning the balancing of demand and supply.

When talking more specifically of electricity markets, the spatial distribution of the network, its interconnections, and the local variability of demand and supply provide an additional element to be considered. Bohn et al. (1984) specified a model to derive optimal pricing not only over time, but also over space, given the network constraint and the different conditions of supply-demand balance. As in the standard approach, they assume a single welfare-maximising public utility, which owns and operates multiple generating plants and sell to independent customers. Demands and supply are both stochastic and they are *spatially located*, flowing in a fixed network subject to stochastic outages (losses). Further, they make the assumption that utility can set and communicate price *instantly*, and can set a different price for each customer location (*node*) at each moment, thus inducing socially optimal behaviour *without* need of rationing. Demands are assumed to be *independent over time* (no cross-price elasticity), so that the model can be solved as a single-period deterministic model.

In this framework, the standard welfare criterion of maximising consumers' plus producers' surplus is constrained to the energy balance and to the network constraints at each location (transmission constraints). The Lagrangian multipliers of the various constraint can be interpreted, as usual, as shadow prices. The solution gives the optimal spot price at each node, and can be described in the following way[10]:

p_i^* = [social cost of additional demand at a general location]
x [1 + incremental losses caused by node *i*]
+ [transmission constraint terms, summed over lines]

The first term refers to the Langragian multiplier associated to the energy balance constraint, and represent the shadow price of an additional unit of demand. This value is the same at each node, and turns out to be the optimum at each consumer location if there are not incremental losses associated to an increase in demand, and no transmission constraint is binding. The second term accounts for different effect on losses of an increase of demand at the various consumer locations, thus charging a higher price to customers whose demand generate a higher marginal loss in the network. Finally, the third term considers the transmission constraints related to the limited physical capacity of the network. Each node can experience congestion, i.e. the constraint can be binding, and in this case the shadow price of an additional unit of transmission capacity will not be 0. The congestion charge at each location is defined as a weighted average of all Lagrangian multipliers: this implies that this component of the locational marginal price can be different from 0 also in a node which does not directly experienced congestion. Potentially, given the network interconnections, it is sufficient to have congestion in a single node to generate positive (or negative) congestion charges at each different node. The same optimisation process is repeated over time (real-time pricing), generating different energy price at each location and in each time-period (e.g. each hour).

5. Competitive electricity markets: wholesale vs. retail prices.

In Italy and in many other countries, electricity markets have been involved in a serious restructuring process, aiming at introducing competition among operators. The previous model of a vertically integrated unit (from generation to retailing) has been reconsidered, since competition can be introduced in the phases of generation and retailing, but not in transmission and distribution services, which are still seen as natural monopolies, due to the network characteristics. Regulators have to find appropriate mechanisms to promote competition ensuring reliability of the service. One possible

[10] For the analytical solution to the problem, see Bohn et al. (1984).

solution directly comes from the theoretical derivation of optimal pricing by Bohn et al. (1984). Current market design often requires electricity to be sold in a *spot wholesale market*, where potential buyers (retailers or final consumers) and sellers (generators) submit their bids for each hour of the day. A centralised system operator observes both demand and supply at each location, and derives real time equilibrium prices as a consequence of the auction, taking into account the network constraints and transmission and distribution costs.

However, the regulation of the electricity market is even more complex. First, bids are generally submitted in two distinct markets: the day-ahead and the real-time (adjustment) market. Moreover, of course, not all customers are able to submit hourly bids in the pool, and there is the need of a *retail sector*. The latter usually charges to final consumers prices which does not directly depend on the spot price fluctuations[11]. Since end users simply do not see the "true" spot price, they can not base their decision on it, and this behaviour reflects into the wholesale demand, which results almost completely unresponsive to price in most power markets. In fact, according to Lafferty et al. (2001), wholesale buyers rarely submit price-sensitive bids; on the contrary, they typically submit bids stating only the quantity to be purchased. Actually, most of them are distribution utilities that have a legal obligation to provide electricity to their customers. Since the latter usually face fixed retail prices, so that they do not have any incentive to respond to hourly wholesale prices, also utilities bids cannot be price-sensitive. It is clear the failure in at least one of the necessary conditions stated by Vickrey (1971) to have spot pricing as a first best solution, because "customers (*utilities)* can *not* respond optimally to price signals"[12].

Summarising, the theoretical optimality of the pricing scheme proposed by Bohn et al. (1984) cannot be directly applied to "competitive electricity markets" because of the existence of a retail sector. As pointed out by Borenstein and Holland (2004), the literature described above has focused entirely on time-varying prices in a *regulated* context. These results "…carries over immediately to a deregulated market only if *all* customers are on real time pricing, but that situation is unlikely to occur in any electricity system in the near future". Therefore, they studied the impact on efficiency of competitive power markets of having some customers on time-invariant pricing. In their framework, a fraction α of the customers pays real time prices and the remaining share (1-α) faces a flat retail rate ($\bar{p}$)[13]. The fraction α is an exogenous number over the interval [0,1], and the aggregate demand is therefore given by:

[11] The characteristics of the different price schedules that can be charged to final consumers will be analysed in the next section.

[12] Also risk neutrality and the absence of transaction costs are strong assumptions.

[13] The fraction of customers on real time pricing is assumed to react optimally to price signals.

$$\tilde{x}_t(p_t, \bar{p}) = \alpha x_t(p_t) + (1-\alpha) x_t(\bar{p}) \qquad [5]$$

The model assumes the following structure of the market:

a) there is perfect competition among generators. Coherently with the previous sections, the cost of installing a unit of daily capacity is β and generators can produce up to the installed capacity with a marginal operating cost equal to *b*;

b) each hour (*t*), generators sell electricity in the wholesale pool market at a price w_t;

c) retail sector is assumed to have no costs other than the wholesale cost of electricity, and firms engage in retail competition.

Competition among retailers forces equilibrium real time price p_t^e to be equal to the wholesale prices w_t. The zero profit condition[14] for the retail sector yields the equilibrium flat rate ($\bar{p}^e$), which is equal to the *demand*-weighted average wholesale price [7].

$$\Pi_{retail} = \sum_t \alpha[x_t(p_t)(p_t - w_t) + (1-\alpha)[x_t(\bar{p})(\bar{p} - w_t)] \qquad [6]$$

$$\bar{p}^e = \sum_t w_t \left\{ x_t(\bar{p}^e) / \sum_t x_t(\bar{p}^e) \right\} \qquad [7]$$

In the wholesale market, the intersection between demand and supply yields the short-run competitive equilibrium:

$$(w_t^e - b)\,\{x_t(w_t^e, \bar{p}^e) - K] = 0 \quad \text{for each period t} \qquad [8]$$

Condition [8] implies that whenever there is enough installed capacity, the wholesale price will be equal to the marginal cost; instead, when demand is higher than K, the wholesale price will increase until the demand/supply balance is achieved. Thus, generators make short-run profits, while in the long the zero profit condition holds:

$$\sum_t (w_t^e - b)\,\{x_t(w_t^e, \bar{p}^e)] = \beta K \qquad [9a]$$

which can be rewritten as[15]:

$$\sum_t (w_t^e - b) = \beta \qquad [9b]$$

As in the classical peak load theory, prices include the capacity payment only at the peak. The inefficiency comes from the retail market, and in particular from the determination of the flat rate. Borenstein and Holland (2003, 2004) demonstrated that a competitive market fails to achieve the second-best optimum given the constraint of having a share of customers paying time-invariant prices. Indeed, if a social planner were to choose the prices p_t^* and ($\bar{p}^*$) that maximise social welfare, in the short run, he would have solved the following optimisation procedure:

$$\max_{p_t, \bar{p}} \sum_t [\tilde{U}(\tilde{x}_t(p, \bar{p}) - b\tilde{x}_t(p, \bar{p})] - \beta K \qquad \text{s.t.} \quad \tilde{x}_t(p, \bar{p}) \le K \quad \text{for all t} \qquad [10]$$

[14] Note that the assumptions on retail sector imply that zero profit condition holds also in the short run (there are no fixed costs).

[15] This is possible because margins are positive only when $x_t(w_t^e, \bar{p}^e) = K$

$$p_t{}^* = b + \lambda_t \qquad [11]$$

$$\bar{p}^* = \frac{\sum_t p_t{}^* \dfrac{dx_t(\bar{p}^*)}{d\bar{p}^*}}{\sum_t \dfrac{dx_t(\bar{p}^*)}{d\bar{p}^*}} \qquad [12]$$

The real-time prices are equal to the marginal cost whenever the capacity constraint is not binding (λ_t is the shadow price of capacity, and is positive only when installed capacity is not enough to face the demand for that period). As to the optimal flat rate, it is the average of the real-time (wholesale) prices weighted by the *slope* of the demand: thus, difference between [12] and [7] comes from the different weights used, and $\bar{p}\,*$ can be higher or lower than $\bar{p}^e$.

In the long run, the second best would be implemented when [10] is maximised also with respect to the amount of capacity K, and yields a further first order condition:

$$\sum_t \lambda_t = \beta \qquad [13]$$

The inefficiency in the long run still comes from the determination of the flat rate and from the comparison between [12] and [7]. At the same time, also the optimal capacity can be either higher or lower than the competitive equilibrium capacity, depending on the relation between equilibrium and socially optimal flat rates: if $\bar{p}* > \bar{p}^e$, then $K^* < K^e$ and vice versa. For example, if we are in a long run competitive equilibrium (so that condition [9b] holds) and $\bar{p}* > \bar{p}^e$, then a regulator may try to improve welfare by increasing the flat retail price. This would reduce demand from flat rate consumers and, since in the short run the wholesale equilibrium price is derived from the supply/demand balance (condition [8]), w_t would decrease in all periods when capacity was fully utilised[16]. Thus, condition [9b] does not hold anymore, because there is excess of capacity: the long run equilibrium would imply therefore a lower amount of capacity.

In conclusion, a regulator with *perfect information* on demand curves performs better than a competitive market, given the constraint of flat rate consumers. From one hand, this gap worsens in a situation of oligopoly among generators. The part of consumers on flat rate is inelastic to changes in wholesale prices; inelastic demand carries with it a higher possibility for the supply-side to exert market power. From the other hand, the quality of information available to the regulator is crucial to perform better than the market. Moreover, Joskow and Tirole (2004) show that the results of

[16] Consider a peak period when there is a problem of excess demand. In the absence of rationing, prices in the wholesale market must raise to reduce the consumption of real time consumers, until the demand/supply balance is obtained. When the flat rate increases, the demand will be lower in all time-periods, and also during peak periods. Then, the problem of excess demand will be mitigated, and a lower wholesale price would be needed to achieve the balance.

inefficiency in a competitive market can be overcome if the retailers are not constrained to offer linear prices, but are allowed to propose two-part tariffs. In their model they also allow for rationing, so I let to the next section a more detailed explanation.

5.1. Rationing

In the literature described above, the problem of having demand which varies over time is solved by means of appropriate pricing schemes. An alternative to this "pure" pricing approach may be rationing supply during high demand states: in this case, a part of the demand cannot be satisfied, but prices paid by the consumers could be considerably lower with respect to real-time rates. Indeed, there are many markets in which rationing behaviour is commonly observed: e.g. restaurants, hotels. Tickets for important events are usually rationed and market clearing prices are not applied, also for a matter of fairness. Gilbert and Klemperer (2000) demonstrated that committing to a fixed price and rationing when there is excess demand may be more profitable than the best market clearing price schedule (even though rationing is inefficient *ex-post*). In particular, they refer to a situation where consumers must incur sunk costs to enter the market. Their result can be applied in the context of *auction theory with endogenous entry*: "when buyers have costs of entering an auction (i.e. sunk investment costs), the seller may wish to precommit to running an inefficient auction (i.e. rationing), to encourage the entry of buyers […] with lower values". When thinking to the wholesale electricity market, there are certainly relevant costs of participation, especially in terms of learning. The introduction of some form of rationing (e.g. by setting a maximum price for each hour) may be an incentive for increasing the number of buyers in the electricity pool, with positive effects for the market liquidity[17].

Joskow and Tirole (2004) argued that rationing of price-insensitive consumers may be optimal if peak periods are infrequent: in this case the peak price tends to infinity and the discrepancy with the fixed price paid by flat rate consumers is too large to make it socially optimal to serve the consumers. Formally, they consider a generalisation of [10], defining a continuum of states of nature $t \in [0,1]$ whose frequency is denoted by f_t, and allowing the social planner to ration demand. In particular, $\psi_t \in [0,1]$ is the share of demand which is served, so $\psi_t < 1$ implies rationing. As in Borenstein and Holland (2003, 2004), the demand is splitted between price-sensitive and price-insensitive

[17] The liquidity of the market is the share of electricity that pass through the pool. Though this share is quite variable among the different world markets (and depends on the features of the market design), it is usually a low value (e.g. for the Italian market it was around 30% during May 2004). Actually, the preference of the majority of customers for signing long-term contracts reflects on these figures.

consumers, so they allow for different values of rationing for real-time ($\hat{\psi}_t$) and flat-rate consumers ($\overline{\psi}_t$)[18].

$$\max_{p_t, \overline{p}, \hat{\psi}_t, \overline{\psi}_t} \int_t f_t[\widetilde{U}(\widetilde{x}_t(p_t, \overline{p}, \hat{\psi}_t, \overline{\psi}_t) - b\widetilde{x}_t(p_t, \overline{p}, \hat{\psi}_t, \overline{\psi}_t)]dt - \beta K$$

$$\text{s.t.} \quad \widetilde{x}_t(\cdot) \le K \quad \text{for all t} \qquad [14]$$

The results of this maximisation imply that price sensitive consumers (facing real time prices) should never be rationed ($\hat{\psi}_t$=1). Instead, for flat-rate consumers:

$$\text{either} \quad \frac{\partial \widetilde{U}_t \big/ \partial \overline{\psi}_t}{\partial \widetilde{x}_t \big/ \partial \overline{\psi}_t} = p_t \qquad \text{or} \quad \overline{\psi}_t = 1 \qquad [15]$$

Thus, in case of rationing, the real time price must be equal to the marginal surplus associated with a unit increase in supply to the (flat-rate) consumers (i.e. the value of lost load, VOLL). To see that rationing can be optimal, suppose that blackouts can be perfectly anticipated (foreseen rolling blackouts). In this case, it is fair to assume that both utility and demand are linear in ψ_t, which implies that VOLL is simply the average gross consumer surplus[19]. Rationing is preferred to market clearing prices mechanisms if and only if the value of lost load is lower then the market clearing price. In the case of only two time-period, rationing should arise only in the peak, and it would be optimal if:

$$U_p(x_p(\overline{p})) < p_p x_p(\overline{p}) \qquad [16]$$

where the subscript p indicates the peak period. If the frequency of peak tends to zero, then peak price goes to infinite, while the optimal flat rate is still bounded; thus the right hand side of [16] is infinitely high and the condition for optimal rationing is verified.

Joskow and Tirole (2004) state that a competitive electricity market *can* achieve the second-best given the constraint of price-insensitive consumers (but whose real time consumption can be measured), given that retailers (or Load Serving Entities, LSEs) are able to offer two-parts tariffs. The condition is that rationing may occur only for price-insensitive consumers and must make use of available generation (i.e. foreseen rolling blackout); furthermore, LSEs can demand any level of rationing they prefer contingent to the real time price[20]. Clearly, these are highly restrictive hypothesis, because actually unforeseen blackouts can occur: in this cases network collapses and there is loss of

[18] In the model, they also allow for more complex technology (production costs and investment costs are different between baseload and peakers), however this aspect can be simplified to our purposes.

[19] Linearity implies $\widetilde{x}_t(\cdot, \psi_t) = \psi_t \breve{x}_t(\cdot)$, thus the derivative with respect to ψ_t yields $\breve{x}_t(\cdot)$. The same reasoning applies to utility, so that $VOLL = \breve{U}(\breve{x}_t(\cdot)) / \breve{x}_t(\cdot)$.

[20] A further limitation of the result above comes from the assumption of homogeneity among consumers (up to a scale factor). Heterogeneity among consumers may result in problems of adverse selection and competitive screening.

available generation. Reliability calls on the need for operating reserves and their optimal definition.

6. Peak load pricing in practice: demand side participation programs

When consumers pay time-invariant rates, wholesale price fluctuations reflecting the supply-demand balance are not passed on to retail customers, and therefore their decisions are independent from the actual system load situation and from the marginal cost of production. In this section, the objective is to describe various alternative to flat rates which have been proposed and implemented in practice.

A *demand-side participation program* can be defined as any possible method used to make the economic incentives of customers more accurately reflect the time-varying wholesale cost of electricity. There are many different possibilities to achieve such a goal of a price-responsive demand. Table 2 provides a list of these methods and describes their capability to give an efficient signal of the real time demand-supply balance. This is clearly related to the possibility of varying the retail price on a short notice. RTP, which implies different retail prices for every hour of the day, varying every day, can achieve this goal almost perfectly, depending on the lag between the price announcement and the price implementation. In its extreme (virtual) application, the real time price for each hour is announced at the beginning of the hour. However, where it has been implemented, the prices for all hours of a day are typically announced on the previous day, with the participants to the program informed via fax or/and internet (for example, on 24 July at 4 o'clock participants receive a fax containing the prices valid on 25 July from midnight to 1 o'clock, from 1 to 2, and so on). The more the lag increases, the more RTP becomes in a certain way similar to TOU, loosing the efficiency in reflecting the true variation in the wholesale market. Thus, a TOU structure entails preset prices based on the average wholesale variation, and for this reason it is not able to capture an unexpected shock.

To summarise, the fundamental difference between TOU and RTP lies in a *static* versus *dynamic*[21] approach to retail pricing. It is also interesting to note that the other methods listed in Table 2 can be viewed either as an improvement of TOU (demand charges that are usually implemented together with TOU, and especially CPP), either as a particular form of CPP. The latter is a sort of a mixed system that uses a TOU static

[21] Here I use the adjective "static" to indicate a preset structure like TOU. Actually, TOU prices can also be periodically adjusted, but this usually happen just a few times a year. At the opposite end, I use "dynamic" to indicate that the adjustment is very frequent, even if it is never continuous.

structure, but adding one more "dynamic" rate that can be called on a short notice to take into account of critical peak hours. Interruptible demand programs and real time demand reduction programs can indeed be viewed also as forms of rationing, even if the participants actually retain an option to continue to consume at a greatly increased price.

Table 2. Demand-side participation programs

	Definition	**Signal of the actual supply/demand balance**
Real Time Pricing (RTP)	Retail electricity prices that fluctuate with the real time wholesale prices	Accurate, depending on the lag time between the price announcement and the price implementation
Time-of-Use Pricing (TOU)	Retail electricity prices varying in a preset way within certain block of time	Approximate, since prices don't capture the price variation within a price block. Moreover, they are based on the average wholesale market variation and adjusted infrequently
Demand Charges	Instrument that allows a portion of the consumer's bill to be calculated on the basis of the consumer's maximum capacity usage	Approximate, since the charge is based on the individual peak and not on the system peak
Critical Peak Pricing (CPP)	System that usually starts with a TOU rate structure, and adds one more rate that applies to critical peak hours, which the system operator can call on short notice	Good, but less accurate than RTP for two reasons: first, the level of prices for the peak hours are preset; second, the number of peak hours that can be called in a year is limited.
Interruptible Demand Programs	System with a basic constant rate structure, with the option for the system operator to cut off supply to some customers.	Since the customers are not actually physically interrupted, but they retain an option to continue to consume at a greatly increased price, these programs can be viewed just as a crude form of CPP.
Real Time Demand-Reduction Programs (DRP)	System where certain customers are eligible to be paid to reduce their consumption at certain times.	Similar to interruptible demand programs

The benefits from allowing dynamic pricing can be shown graphically referring again to the basic peak load model in Figure 1. There are only a peak and an off-peak demand and that market is competitive. Thus, supposing an installed capacity of K, than, if time-varying rates are allowed, the prices will be p_p and p_{op} during the peak and the off-peak period respectively. K represent an optimal capacity and there is no incentive to invest more.

If the price is constrained to be at the unique rate P_F, then the effects will be the following (Borenstein, 2003):

- an inefficient decrease in the off-peak consumption, causing a deadweight loss;

- a demand exceeding the supply at the peak rate, involving the need of some sort of rationing.

This second aspect would produce an incentive for firms to over-invest in capacity. Since in peak period they must sell at p_F and they cannot charge an higher price, there is an incentive to build new capacity to meet the additional demand. The author emphasises the role of time varying prices, that encourage customers to consume less in peak periods avoiding this excess of capacity. Moreover, if the wholesale market is not competitive, with fixed retail price it is much more profitable for the wholesale seller to exercise *market power*. In fact, a raise in the wholesale price has no short-run impact on sales since end-use customers do not see a change in their bill.

Now suppose that a TOU structure is used and consider the effect on the simple model described above. In this case, we have an improvement because there will be two rates, p_o^F and p_p^F, which however can only approximate the competitive unconstrained prices p_o and p_p, since they are fixed ex-ante. In the real world, since there are not only two time periods, and both peak demand and especially time when peaks occur are difficult to predict, the approximation can be very inaccurate with respect to RTP. An empirical investigation for the summer of 2000 in California has shown that less than 20% of the variation in the wholesale market could have been reflected in a TOU structure, even setting the TOU prices ex-post[22].

Though RTP is more efficient, TOU have been more widely used and accepted, in part because it is easier and less costly to implement. RTP benefits must be high enough to justify investments in metering, and needs efficient systems of communications. However technology is evolving fast, and can support the implementation of RTP at least in three directions: first, making available sophisticated metering technology at a reasonable cost; second, simplifying communication thanks to the internet; third,

[22] This investigation was based on a regression of the hourly wholesale price on dummy variables for each of the TOU periods, and the R-squared of such a regression provides the share of price variation captured by using TOU periods rather than a single constant price (Borenstein, 2003).

enhancing the ability to respond to frequent retail price signals, that sometimes could be achieved without the human intervention thanks to the use of “smart” energy management systems. Borenstein et al. (2002) states that the cost of this investment may not be feasible for very small users, but would be certainly desirable for large users.

A part from the technological barriers, there are also cultural and regulatory barriers to RTP (Yoshimura, 2003). For example, it is a common belief that having electricity is a basic right, and that prices should be time invariant. Even if time-variant prices would produce savings. Moreover, policies usually support this belief, requiring the utilities to offer time invariant retail electricity prices. According to Borenstein et al. (2002) the concerns about RTP typically involve three types of issues: the customer price risk, equity concerns and mandatory versus voluntary programs.

a) Hedging against the risk

Because the real-time or the day-ahead price of electricity is highly volatile, customers are diffident towards RTP, for the risk of paying drastically increased prices during certain hours. This involve the need to create some form of insurance for the consumers, by purchasing some power on long term contracts in order to give a certain stability to their monthly bills. One approach is to implement a two-part RTP program with a Customer Baseline Load (CBL), that allows consumers to buy a certain amount of power according to standard TOU rates, while they face real-time rates when their consumption increase over a certain predefined level. However this raises difficulties on the definition of the CBL. Rather than assigning a certain baseline level, it seems more appropriate allowing the customer to purchase a baseline (with a forward contract) to hedge as much he desires. The fact that incremental consumption decision are still subject to RTP ensures strong incentives to conserve at peak times.

b) Equity issues

Maybe the most important diffidence against RTP is the fact that such tariffs would necessarily involve an arbitrary redistribution among different types of customers. Of course, the most flexible consumers and those that usually tend to have a smoother consumption will be the first ones to gain from RTP, while customers with more “peaky” demand, unwilling to switch their consumption, will pay a high share of their power at the more expensive rates. However, the latter could expect to gain from positive externalities coming from the reduction of peak consumption by the most flexible consumers. In fact, lower peak demands mean less investment in excess capacity and therefore lower payments to the generators in the wholesale market. This is even more considerable if we consider the the total capacity is built on the basis of the system peak, but in order to minimise the risk of blackouts there are of course reserve requirements (usually set between 10 and 20 per cent of the peak demand). Price responsive demand will not only imply a lower system peak, but also a reasonable lower percentage of reserve requirement. This is because the increase in peak price will at least partially absorb an unexpected system shortage. Moreover, RTP reduces the ability of sellers to exercise market power. The point is to understand the extent of these benefits in order to evaluate the feasibility of the program.

c) Mandatory or voluntary programs

If the gains from dynamic pricing depends crucially on the customer load curves, then one of the possibilities is to implement a voluntary program. This would allow the most inelastic users to stay at fixed rates. However, a voluntary approach can give raise to a problem of adverse selection, if its implementation generate a cross subsidisation from RTP users to the others. This could happen because the retailer will see a decrease in its revenues (since users will choose RTP only if they can save money). To keep its revenue at the same level he will decide to charge an adder on RTP, in order to equalise the average price between participants and non-participants. But this will clearly undermine the incentives to join the program. In order to be successful, a non-compulsory RTP program must have a commitment of no cross subsidisation.

7. Conclusions and future research

The idea of peak-load pricing originated in the context of regulated public utilities industries, motivated by the necessity of choosing the optimal level of capacity and covering the full social cost of providing the service. The deterministic model in Steiner (1957) has been progressively extended to consider a more complex and realistic framework. In particular, the uncertainty on demand and supply realisations implies a problem of reliability of the service, giving a further motivation for time-differentiated tariffs. In this context, the latter can be used by the regulator as instruments of demand management, to reduce the probability of blackouts. The problem of optimal pricing takes a *dynamic* aspect, since it would be necessary to adjust tariffs continuously, to take into account of the stochastic variations in the demand-supply balance. These considerations give raise to the concept of real time pricing, and in parallel, to the idea of interruptible tariffs. When considering the specific electrical network and the transmission constraints, price differentiation can be used to solve congestion problems, not only over time, but also over space, as in Bohn et al. (1984).

In many restructured electricity markets, and also in Italy, the idea of spot pricing has somehow been implemented introducing the wholesale power exchange market. A centralised system operator observes hourly bids from buyers and sellers, and derives hourly equilibrium prices as a consequence of the auction, taking into account the network constraints and transmission and distribution costs. This design has been studied with the aim of promoting competition among operators in the phases of generation and retailing. Apart from any consideration on the desirability of competitive markets, the theoretical efficiency of a spot wholesale market as an instrument of demand management can be seriously undermined if wholesale buyers are typically insensitive to hourly price variations; this is usually the case considering the structure of the retail market, where final customers generally face prices which are independent from the hourly wholesale price fluctuations. Indeed, the traditional literature on real-time pricing refers to a regulated public utility, and its results cannot be applied directly to a deregulated market unless all final customers are on real time pricing (or, in other word, if there is not a retail market). This is clearly not the case in any electricity markets, and, moreover, the existence of a retail market can be justified in terms of risk aversion, equity issues, and in general from the fact that not all consumers are able to respond to dynamic price signals.

The relation between wholesale and retail prices is crucial to the efficiency of the market, and therefore the restructured electricity market has provided a new context for theoretical works concerning time-varying prices (Borenstein and Holland, 2003 e 2004; Joskow and Tirole, 2004). A first general question is related to the (de)regulation of retail market: is it better to promote retail competition or to regulate retail tariffs? Joskow and Tirole (2004) stated that a competitive retail market can perform as well as a regulator with perfect information, even with the constraint of having a share of price-insensitive consumers, but whose real time consumption can be measured, given that retailers are able to offer two-parts tariffs and under some restrictive conditions on efficient rationing. Another question is more strictly related to the desirability of time-varying rates, and to the real level of competition among operators. In particular, an inelastic wholesale demand may favour market power behaviour of generators, if they were not in perfect competition. In this sense, retail time-differentiated rates assume a further role as a mean for promoting demand responsiveness in the spot wholesale market.

Generally speaking, the available literature relies on simplifying assumptions over the characteristics and the behaviour of consumers. Perhaps the most intriguing area of future research lies in the explicit modelling of consumers heterogeneity, both in terms of load profiles and price responsive attitude. No research works have been done to explore formally the conditions under which a customer can be willing to switch from a flat rate towards a dynamic rate, while retailers being willing to offer time-varying prices at the same time. This is relevant because, if time-varying rates have to be voluntary, an adverse selection problem may arise.

Summarising, the question of optimal pricing in the electricity sector is far from being solved. In the theoretical literature, we have found basically three motivation in favour of time-differentiated pricing: a) the question of optimal capacity and the efficient use of resources (peak load pricing in general); b) being an instrument of demand management related to the problem of stochastic demand-supply balance and reliability (real time pricing or interruptible tariffs); c) reducing potential market power behaviour of generators in the wholesale spot market (demand side participation programs). From the other hand, however, especially when referring to dynamic pricing, one should consider the effectiveness of price responsive behaviour that can be induced in the final consumers, together with risk aversion preferences and equity issues. Indeed, dynamic pricing may produce benefits only if it gives to the consumer the possibility to perceive the price signal (without relevant costs), thus in the presence of enhanced communication systems and/or automatic energy management systems.

REFERENCES

Bohn R., Caramanis M. and Schweppe R. (1984), “Optimal Pricing in Electrical Networks Over Space and Time”, *The Rand Journal of Economics*, 15(3), 360-376.

Boiteux, M. (1949), “La Tarification des Demanded en Point: Application de la Théorie de la Vente au Cout Marginal.”, *Revue Generale de l’Electricité* 58, 321-40; translated as “Peak Load Pricing.”, *Journal of Business* 33(2), (1960), 157-79.

Borenstein S. (2003), "Time-Varying Retail Electricity Prices: Theory and Practice," in Griffin and Puller, eds., *Electricity Deregulation*: *from Where to Here*, Chicago, University of Chicago Press.

Borenstein S. and Holland S.P. (2003), *Investment Efficiency in Competitive Electricity Markets With and Without Time-Varying Retail Prices*. Center for the Study of Energy Markets. Working Paper CSEMWP-106R.

Borenstein S. and Holland S.P. (2004), “On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices”, forthcoming on *The Rand Journal of Economics*

Borenstein S., Jaske M. and Rosenfeld A. (2002), *Dynamic Pricing, Advanced Metering, and Demand Response in Electricity Markets*. Center for the Study of Energy Markets. Working Paper CSEMWP-105

Brown G.Jr. and Johnson M.B. (1969), “Public Utility Pricing and Output Under Risk”, *American Economic Review*, 59(1), 119-129.

Bye R.T. (1926), “The Nature of Fundamental Elements of Costs”, *Quarterly Journal of Economics*, 41, 30-63.

Crew M.A., Fernando C.S. and Kleindorfer P.R. (1995), “The Theory of Peak-Load Pricing: A Survey”, *Journal of Regulatory Economics*, 8, 215-248.

Crew M.A. and Kleindorfer P.R. (1976), “Peak-Load Pricing with a Diverse Technology”, *Bell Journal of Economics*, 7, 207-231.

Gilbert R.J. and Klemperer P. (2000), “An Equilibrium Theory of Rationing”, *The Rand Journal of Economics*, 31(1), 1-21.

Houthakker, H.S. (1951), "Electricity Tariffs in Theory and Practice", *Economic Journal*, 61, 1-25

Ilic M., Black J.W., Fumagalli E., Visudhiphan P. and Watz J.L. (2001), *Understanding Demand: The Missing Link in Efficient Electricity Markets*, Energy Laboratory Publication, MIT, EL 01-014WP.

Joskow P. and Tirole J. (2004), *Reliability and Competitive Electricity Markets*, presented at IDEI-CEPR conference on "Competition and Coordination in the Electricity Industry", Toulouse, January 2004.

Kleindorfer P.R. and Fernando C.S. (1993), "Peak-Load Pricing and Reliability under Uncertainty", *Journal of Regulatory Economics*, 5(3), 317-336.

Lafferty R., Hunger D., Ballard J., Mahrenholz G., Mead D. and Bandera D. (2002), *Demand Responsiveness in Electricity Markets*, presented at FERC-DOE Demand Response Conference, February 2002.

Shy O. (2001), *Dynamic Peak-Load Pricing*, mimeo *econ.haifa.ac.il/~ozshy/peak37.pdf*

Steiner P.O. (1957), "Peak Loads and Efficient Pricing", *Quarterly Journal of Economics*, 71, 585-610.

Stoft S. (2002), *Power System Economics: Designing Markets for Electricity*, IEEE Press, Wiley-Interscience.

Vickrey W.S. (1971), "Responsive Pricing of Public Utility Services", *Bell Journal of Economics*, 2, 337-346.

Visshler M.L. (1973), "Welfare-Maximizing Price and Output with Stochastic Demand: Comment", *American Economic Review*, 63(1), 224-229.

Williamson, O. E. (1966), "Peak-Load Pricing and Optimal Capacity under Indivisibility Constraints.", *American Economic Review*, 66(4), 589-97.

Yoshimura H. (2003), *Making Demand Response Work in New England*, presented at the Northeast Energy and Commerce Association, January 2003.