Trading Hubs Construction for Electricity Markets

Pavel A. Borisovsky[1], Anton V. Eremeev[2], Egor B. Grinkevich[3], Sergey A. Klokov[4], and Andrey V. Vinnikov[5]

[1] Omsk State Technical University. 11 Prospect Mira, 644050, Omsk, Russia
   `borisovski@mail.ru`
[2] Omsk Branch of Sobolev Institute of Mathematics SB RAS. 13 Pevtsov St., 644099, Omsk, Russia `eremeev@ofim.oscsbras.ru`
[3] Administrator of Trade System for United Energy System of Russia. 12 Krasnopresnenskaya Naberezhnaya, 123610, Moscow, Russia
   `geb@rosenergo.com`
[4] Omsk Branch of Sobolev Institute of Mathematics SB RAS. 13 Pevtsov St., 644099, Omsk, Russia `klokov@ofim.oscsbras.ru`
[5] Administrator of Trade System for United Energy System of Russia, Joint Institute for Nuclear Research. 12 Krasnopresnenskaya Naberezhnaya, 123610, Moscow, Russia `vinnikov@rosenergo.com`

Abstract

In this chapter, we consider a problem of constructing trading hubs in the structure of the electricity wholesale markets. The nodes of a trading hub are used to calculate a reference price that can be employed by the market participants for different types of hedging. The need for such a reference price is due to considerable variability of energy prices at different nodes of the electricity grid at different periods of time. The hubs construction is viewed as a mathematical programming problem here. We discuss its connections with clustering problems, consider the heuristic algorithms of solution and indicate some complexity issues. The performance of algorithms is illustrated on the real-life data.

Key words: Electricity Market, Trading Hub, Mathematical Programming, Clustering, Genetic Algorithm, Complexity

1 Introduction

In the modern electricity spot markets the price is not unique, it varies from one node of the power grid to another and it also depends on time. The market participants in this situation are interested in one or several reference prices to hedge the price risks and to settle the forward contracts. These reference

prices can be calculated by taking an average of the energy prices in a number of nodes with the most typical price dynamics in the given region. A set of such nodes with a specific formula for computing the average is called a *trading hub*. For short, in what follows, we will use the term *hub*.

Large electricity markets, such as PJM Interconnection (USA), Midwest ISO (USA), United Energy System (Russia) and others, provide a number of hubs. In this case, each buyer or seller prefers the hub approximating the most closely the nodal price of this participant. The hubs in electricity markets have some similarity with the hubs in oil and gas markets, but each of these commodities has unique features which require relevant trading instruments [2].

Successfully functioning hubs contribute to emergence of *derivatives*, the financial instruments (contracts) that do not represent ownership rights in any asset but, rather, derive their value from the value of the underlying commodity. The derivatives may serve as efficient tools for isolating financial risk and hedging to reduce exposure to risk [6]. The hubs also contribute to the success of electronic trading systems providing the aggregated data on price dynamics over the system. In the case of electricity markets, the hub price is usually defined as a simple or a weighted average of the nodal prices over the nodes comprising the hub. Due to this reason the hub price is less volatile than the prices at individual nodes. This feature is of particular importance because the liquidity of futures contracts depends significantly on predictability of the price of the underlying commodity (liquidity here means large volume of trade operations and easiness to find a contracting party).

To define a hub, it is sufficient to select a set of nodes and the weights to average the prices over these nodes. In the present chapter, we will consider this task as an optimization problem, keeping in mind that usually the optimization is just one of the steps in the decision making process of designing a hub. This process in practice involves a lot of negotiations between the market participants and administration, so that the human expertise often plays an essential role. In some cases it has been proposed to define the hubs without any optimization, e.g. the hubs may consist exclusively of generation nodes, grouped on the regional basis, with weights equal to the historical volume of the generation or the installed capacity, but such approach is not always applicable. One of the promising statistical approaches to hubs construction is based on the Principal Component Analysis [4] but detailed presentation of this method is outside of the scope of this chapter.

The trading hubs construction problems considered below have similar terminology to the *hubs location problems* [3], however, these classes of problems are different. The hubs location problems are mainly motivated by the applications where certain elements of a system are actually connected via hubs, while in our case the trading hubs are purely virtual constructions and no physical connections are associated with them. Also, in the hubs location problems there is no equivalent of the hub price, which plays an important role in our case.

The remaining part of the chapter is organized as follows. In Section 2, we discuss the motivation for the Hubs Construction Problems, the ways of using the hubs in the electricity markets and the properties demanded from them. Here we also provide a brief review of hubs implemented in some electricity markets. The criteria and constraints, formulated in Section 2, are converted into mathematical programming problems in Section 3 and some basic properties of these problems are discussed here. The Hub Construction Problems often turn out to be large scale non-convex optimization problems, which makes it relevant to look for appropriate heuristics to solve them. Some of these heuristics are presented and evaluated in Sections 4 and 5. Section 6 contains the conclusions.

2 Hedging in the Electricity Markets and Hubs Usage

2.1 Price Volatility

The trading hubs construction problems appear in the context of the modern electricity markets based on the *Locational Marginal Pricing* (LMP). LMP is a mechanism for using market-based pricing for managing transmission congestion and thermal losses in the electricity grid. The energy prices at different locations vary due to transmission congestion, which prevents relatively low-price generation from meeting the loads beyond a certain neighborhood. If not the congestion and transmission losses, the energy price would be uniform all over transmission grid. The market clearing LMP price is determined by an *Independent System Operator* (ISO) on the basis of solution to a mathematical programming problem, known as the *economic dispatch problem*. The LMP price at any node is taken to be a Lagrange multiplier of the power flow balance constraint associated with this node. The details of this approach can be found, e.g. in [5, 16, 26, 30]. Computation of LMP prices requires that all market players submit to an ISO their bids for generation and load. If the price and dispatch schedule computation takes place a day before dispatch, this is called the *Day-Ahead Market*. Additionally, an ISO may support other similar markets scheduled at later time, e.g. an Hour-Ahead Market and the Real-Time Market.

Due to the difficulty to store electric energy for significant time and due to high variability of demand for this good during a day, the LMP price of electricity is highly volatile. An example of price behavior can be seen in Fig. 1. This figure contains the Real-Time data of 13 Feb 2007 obtained from the web site of PJM system operator `http://www.pjm.com`. The LMP prices of many energy markets have a strong dependence on the geographical location and the grid topology [27]. This is why in many cases it is important to establish regional hubs, defining reference prices that closely approximate the cost of energy in the area and may be used for hedging (compare the graphs of PJM Eastern hub and node CARKSVI in Fig. 1).
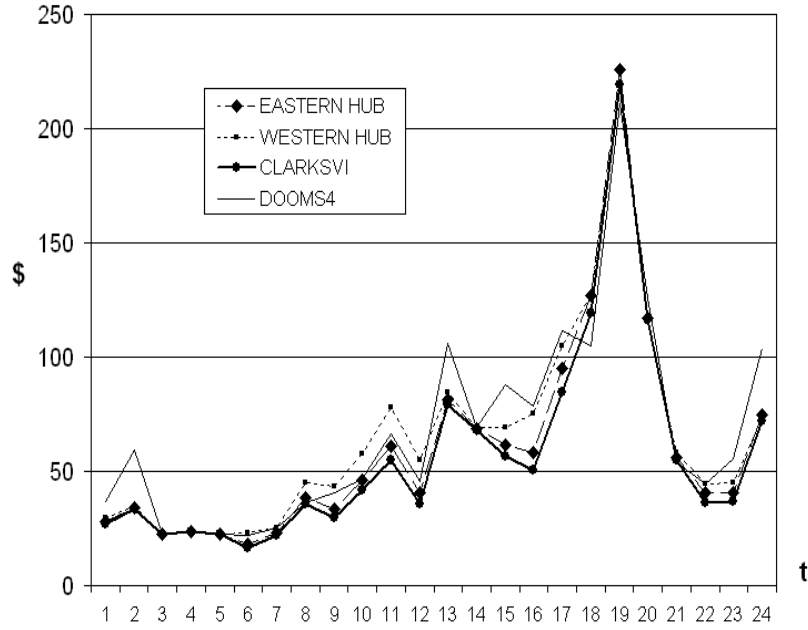
Fig. 1. Real-Time LMP prices at two nodes and two hubs in PJM Interconnection as of 13 Feb 2007

2.2 Basic Hedging Strategies and Hubs Usage

Hedging by Means of Futures Contracts.

Deliveries in the futures market are organized in physical or financial form. The first, physical delivery assumes that the seller at the maturity must hold the specified in the contract quantity of good at the specified warehouse. The seller then sends delivery call to the buyer who transfers money at the price specified in the contract and the seller transfers the right of possession to the buyer. The essence of the contract is its price. Fixing the price when signing the contract allows both the buyer and the seller to secure their cash flow for the future.

The seller of the contract can either deliver the good to the warehouse or buy the good in the warehouse from an agent (third party) at the spot price, while receiving from the buyer the price set in the contract. This observation motivates the second, financial form of delivery.

The financial form of delivery implies that entering into a futures contract at price C_f would yield for the seller the amount of money M_s depending on the spot price C:

$$M_s = C_f - C \,. \tag{1}$$

For the buyer, the result M_b is just the opposite:

$$M_b = C - C_f \,. \tag{2}$$

So, financially the futures contract results for the seller in:

- receiving the difference between the contract and the spot prices when the contract price is higher than the spot price;
- paying the difference between the spot and the contract prices when the contract price is lower than the spot price.

Since electricity can not be stored, the physical settlement would imply that the seller physically delivers electricity at the rate specified in the contract while the buyer transfers money to the seller's account at the specified price. Nevertheless, in the case of electricity, the usage of common electric grid is unavoidable and, therefore, the physical delivery of electricity involves a lot more detailed coordination of actions of all participants compared to physical delivery of other products. For this reason, the more flexible financial form of delivery is widely accepted in the electricity markets. The delivery period is usually one month, the amount is 1 MWh each hour. Also, contracts of 1 week, 1 quarter and 1 year duration are traded at some markets (Nord Pool, EEX).

Hedging by Means of Financial Transmission Rights (FTRs).

The FTRs are hedging instruments, designed for compensation of price difference between the nodes separated by congested transmission lines in the electricity spot markets with locational marginal pricing [14, 16, 26] (e.g. Day Ahead or Real Time markets). The FTR contracts can be defined from any node to any other node. If the hub price is defined as the simple average or a weighted average of a set of nodal prices, then the FTRs may be defined between any node and a hub, or between two hubs. The market of node-to-node FTRs may be quite illiquid due to a large number of nodes in the system. Liquidity can be increased by usage of hubs because an FTR from node 1 to node 2 is decomposable into two FTRs: from node 1 to hub and from hub to node 2 [16]. Traders may obtain the FTRs to cover the basis risk between their own node and the hub or between two different hubs. If the nodal price of a participant is sufficiently close to the price of some hub, the basis risk from node to hub may be quite small with most of the basis risk being between hubs.

In some markets, organized according with the Standard Market Design principles, the rules of the Day-Ahead market allow to post *virtual* demand/supply bids, i.e. the bids for buying or selling the energy, not supported by real generation capacity or load (e.g. New England Pool in the USA). In such a case, a significant amount of virtual demand/supply may be concentrated in a hub and the latter can be viewed as a trading point with risk-hedging potential [24].

The spot price in the electricity markets is significantly volatile even on the daily scale, therefore, when hubs are designed to be used for FTR transactions

or for virtual bidding, it is important to take into account the hourly prices in historical data. Alternatively, if the hub is designed only to be used for one-week, one-month or annual futures, then the statistical data may be averaged over these time periods to simplify the analysis and to design the hub more adequate to the market tools in use.

The Hedge Ratio.

Here we describe the most widely used method of hedging by standard contracts utilizing the so-called *hedge ratio*. This method is well-known among risk managers and it is based on easily calculable and understandable quadratic distance measure between the prices.

Consider a producer selling each hour 1 MWh of electricity on the spot market at price c_{it}, where t denotes hour and i denotes the node of the electric grid to which the producer belongs. The value c_{it} is a random number.

Since the producer sells 1 MWh at price c_{it}, the amount of money $M_{it}^{(s)}$ he gets from the spot market in that hour is

$$M_{it}^{(s)} = c_{it} . \tag{3}$$

The producer is going to hedge the position in the spot market by entering into a financially settled futures contract at price C_f. The settlement of the futures contract is defined relatively to the spot price c_{Ht} in the hub H meaning that according to equation (1) his participation in the futures market results in the value

$$M_{it}^{(f)} = C_f - c_{Ht} \tag{4}$$

per each 1 MWh of the delivery. Note that c_{Ht} is a random value as well.

Suppose the producer located in node i sells h_i futures contracts (h_i may be greater or less than 1). In this case the producer receives a total amount M_{it}

$$M_{it} = M_{it}^{(s)} + h_i M_{it}^{(f)} = c_{it} + h_i \left(C_f - c_{Ht} \right) . \tag{5}$$

For efficient hedging the producer aims to minimize the variance of M_{it} [29]:

$$\text{Min} \quad \sigma^2(M_{it}). \tag{6}$$

The only possible parameter the hedger can change is the amount h of futures contracts that he sells to hedge a unit of sold good in the spot market. Hence, (6) is transformed into

$$\frac{d\sigma^2(M_{it})}{dh_i} = 0. \tag{7}$$

Expanding the variance of the sum and respecting that C_f is not random we obtain

$$\frac{d}{dh_i} \left(\sigma^2(c_{it}) + h_i^2 \sigma^2(c_{Ht}) - 2\rho(c_{it}, c_{Ht}) h_i \sigma(c_{it}) \sigma(c_{Ht}) \right) = 0, \tag{8}$$

$$h_i \sigma(c_{Ht}) - \rho(c_{it}, c_{Ht}) \sigma(c_{it}) = 0, \tag{9}$$

$$h_i = \rho(c_{it}, c_{Ht}) \frac{\sigma(c_{it})}{\sigma(c_{Ht})}, \tag{10}$$

where $\rho(\cdot, \cdot)$ denotes the correlation.

Substituting the value of h_i (known as the *hedge ratio*) into equation (5) and calculating its variance we have:

$$\sigma^2(M_{i,t}) = \left(1 - \rho^2(c_{it}, c_{Ht})\right) \sigma^2(c_{it}), \tag{11}$$

i.e. an optimal hedge ratio decreases the variance of the initial position by factor of $1 - \rho^2(c_{it}, c_{Ht})$. Hence the standard criterion of selecting a good hub for a given node would be to take the one with the maximal correlation with the nodal price.

Unfortunately, reliable estimation of correlation and variance for a given node and given hour is practically impossible due to volatile behavior of electricity prices [29]. One may take the historical prices data in order to calculate the hedge ratio as discussed above, but it will be only a rough estimate.

For the futures contracts with delivery periods of one month duration, the index t in the above relations could be replaced by the month index τ. In such a case, if μ_τ denotes the set of hours belonging to month τ, instead of the values c_{it} one would use

$$c_{i\tau} = \frac{\sum\limits_{t \in \mu_\tau} c_{it} A_{it}}{\sum\limits_{t \in \mu_\tau} A_{it}}, \tag{12}$$

where A_{it} is the amount of electricity sold in the node i at hour t. A discussion of positive and negative factors of averaging with time-dependent weights A_{it}, as compared to the simple average, can be found in [7]. In any case, replacement of the hourly prices by average prices over certain time periods reduces the amount of input data for the decision support system, simplifying the analysis. At the same time, this approach reduces the amount of useful information at the input, e.g. the distinct behavior of on-peak and off-peak prices can not be seen after such an aggregation. With this in mind, we will usually talk about time t indexed in hours, unless otherwise is stated.

It is clear that a large number of hubs would allow to find an appropriate hub for any node. However, large number of futures can not be liquid simultaneously. At most of the exchanges, the number of different liquid futures is small. Usually there exists one lead contract attracting most of liquidity and 3-5 supplementary contracts. Unfortunately, it is difficult to evaluate quantitatively the liquidity of each hub in a given collection of hubs in advance, because this property depends on many organizational factors and strategies of market participants. Hence, the upper bound on the number of hubs is often used as a simplified liquidity requirement.

The necessary number of hubs m may be evaluated by means of the Principal Component Analysis [4]. Usage of this method in hubs design is established

on the basis of assumption that each node i is hedging in all existing hubs with the hedge ratios minimizing (6) w.r.t. each hub. Each eigenvector (principal component) of the sample correlation matrix between all nodal prices can, theoretically, define a hub. The greater its eigenvalue, the greater amount of the total variance of the nodal prices it carries. A decision about the number of principal components, that are practically significant, can be made e.g. using Kaiser's or Cattell's criteria [35].

Another constraint, which is also connected with hub liquidity is a lower bound on the number of nodes in a hub, when the hub price is computed as the simple average of the nodal prices. This constraint ensures stability of the hub price under minor modifications of the grid (permanent exclusion or temporary outage of nodes due to planned repair or unexpected breakdown). In general, taking average over a larger number of nodes usually decreases the variance of hub price, which makes it more predictable for the traders, increasing the liquidity. The lower bound on the number of hub nodes, equal to 50 or 100 appears to be sufficient (see the examples in Section 2.3).

Note that when the hub price is computed as a weighted average, the nodal weight may be chosen arbitrary close to zero, even though, formally, this node is used for calculation of the hub price. This makes it meaningless to impose the lower bound on the number of nodes when the hub price is computed as a weighted average. There are some ways to modify this condition for the weighted case as well, but for simplicity we will consider only the unweighted case here.

To sum up, informally, the Hubs Construction Problem consists in finding a sufficiently small number of large hubs which would explain as much as possible the price dynamics in most of the nodes of the electric grid. The ways to formally state this problem will be addressed in Section 3.

2.3 Hubs Design of Some Existing Markets

In this section, we briefly survey several cases of hubs design in large-scale electricity markets based on locational marginal pricing.

Midwest ISO

The system operator Midwest ISO manages one of the largest electricity markets in the USA. The grid of Midwest ISO consists of more than 30 000 nodes and the LMPs are computed for about 1 500 nodes. The overall installed capacity of Midwest ISO generation is near 150 GW. The electricity market is organized according to the Standard Market Design principles [33]. There are 4 trading hubs in this market: Cinergy, Michigan, Illinois and Minnesota consisting of about 330, 260, 150 and 170 nodes, respectively. These hubs were chosen by LECG, LLC and the Midwest ISO in conjunction with the Trading Hubs Task Force in year 2003 [10].

One of the main requirements in the design of the Midwest ISO hubs was that the hub price should move consistently with the prices in the corresponding target region and most of locations in the target region are "close" to the hub in terms of price difference. Also, it was required that it should be unlikely for a significant portion of the trading hub to be lost from service. The volatility of hub price had to be low, implying that plausible patterns of transmission congestion and individual transmission outages should not cause the trading hub price to substantially diverge from prices in the target region. The trading hub definitions had to be fixed i.e. once a hub was defined, the set of hub nodes and their weights are not changed.

In view of these requirements, the optimization model with given number of hubs was applied (see Section 3.1 below). The input data consisted of the nodal prices for each 2-hour period of the year in 1290 nodes. A straightforward solution of this problem by means of commercial optimization packages was impossible because of its high dimensionality. For this reason the problem formulation was simplified by setting the weights of all nodes equal to 1. This simplification allowed to apply the standard statistical clustering algorithm H-means (see Section 4 below) to form 30 clusters of nodes. Inspection of these clusters showed several sufficiently large candidates with relatively low distance between the hub price and the nodal prices in their target regions. Some of the outliers were manually excluded from these clusters on the basis of the scatter plot analysis and evaluation of the experts. The subsequent analysis consisted in comparison of the hub prices with the electricity price in 37 load area zones of major market participants to determine the cluster best fit for each of them.

PJM Interconnection

operates one of the largest wholesale electricity markets in the world. The overall installed capacity of Midwest ISO generation is near 160 GW, the number of market participants is more than 350. The market structure complies with the Standard Market Design. There are two actively traded hubs: Western (near 110 nodes) and Eastern (near 240 nodes), 8 localized hubs: AEP Gen, AEP Dayton, Chicago Gen, Chicago, Domnion, Northern Illinois, New Jersey, Ohio and an interface hub Western Interface.

The hub price is computed as a weighted average of the real-time LMP prices with a fixed set of the nodal weights. The largest volume of trade is concentrated in PJM-Western hub, due to its stability to the influence of system constraints and its location between large load areas and areas of generation. The PJM-Western monthly futures are traded at NYMEX stock exchange for on-peak and off-peak hours (see `http://www.nymex.com/JM_desc.aspx`).

To determine the composition of the PJM-Western Hub, the nodal prices were analyzed under various historic transmission constrained conditions [25]. The standard cluster analysis tools were used to determine candidate clusters of nodes that respond in a similar way under many different transmission

constraints. For each of these clusters, an optimization problem was solved to determine the node weights that minimize the distance between the hub price and the energy price in the subregion for which the hub is targeted. Originally, in year 1998 PJM-Western Hub consisted of nodes selected from PEPCO, BGE, Penelec and MetEd zones. Later this hub went through some changes, with addition of nodes from APS, ComEd, AEP, Dayton, Dominion, and Duquesne and RECO zones.

ISO New England

system operator is responsible for New England's bulk power generation and transmission system with an installed capacity of 32 GW and more than 200 market participants. The market is organized according with the Standard Market Design. The power grid of ISO New England has only one hub, NEPOOL Hub, allocated between the areas of prevailing generation at North and West and the areas of prevailing consumption at Connecticut and Northeast Massachusetts/Boston. The hub price is the simple average of the nodal prices at 32 nodes. The choice of these nodes was based on statistical analysis, using simulated nodal prices [13].

Electricity Market of United Energy System of Russia

Administrator of Trade System for United Energy System of Russia operates a wholesale electricity market with an overall installed capacity near 200 GW and 200 market participants, more than 100 of them are generators. The Day-Ahead Market is based on the locational marginal pricing, where the hourly prices on electricity are computed for more than 7000 nodes. The mathematical models of the current market, which is functioning since September 2006, and its two-sector predecessor are described in [5]. There are four hubs in the European zone of Russia: Center-Europe, Center-West, Volga and Urals. The other zone is located in Siberia, it has two hubs: Kuznetsk Basin and Krasnoyarsk – see `http://www.np-ats.ru`. The sizes of smaller hubs are close to 50 nodes and the size of the largest hub Center-West is over 300 nodes. The hub prices are computed as simple average of the day-ahead locational marginal prices. All hubs consist of high voltage nodes (not less than 220 kV), which ensures that local congestions and grid modifications do not influence the hub price a lot. The sets of hub nodes in the European zone were chosen using the H-means clustering algorithm with subsequent expertise. The Principal Component Analysis, applied to the European zone indicates that the largest eigenvalue corresponds to the average electricity price in this zone. The first eigenvalue greatly exceeds all other eigenvalues and there are 4 other principle components of significant value, which is consistent with the number of existing hubs.

3 Problem Formulations

In this section, we discuss the mathematical formulations of the Hub Construction Problems, taking into account the criteria and constraints considered in Section 2.2 above.

As it was mentioned before, the trading hubs should be constructed, taking into account the historical data of the nodal prices, preferably, covering a whole preceding year or several years. If the hubs are designed for monthly futures, without separation of on-peak and off-peak futures, then the historical data may be aggregated into one-month elementary periods. If the hubs are also aimed to be used for different types of futures, for virtual bidding at the Day-Ahead market or for the FTR contracts, then such time aggregation is inappropriate. It is important that the input data represent the price dynamics in all seasons and, if the hourly prices are not averaged over one-week or one-month period, it is also important that the data reflects different modes of the system: with congestions and without them, on-peak and off-peak hours, working days and weekends. For terminological convenience, we will usually call the elementary historical time intervals "hours" and denote them by t. The number of elementary time intervals in the historical data will be denoted by T.

Here we assume that the distance measure between the hub price and the nodal price is computed as the sum of squared differences over all hours t, $t = 1, \ldots, T$, i.e. the squared Euclidean distance in T-dimensional space. If appropriate, the Euclidean distance may be substituted by some other standard metric, or by the observed variance or observed correlation (the latter should be maximized) as it was discussed in Section 2.2. The sum of squared differences appears to be the most widely used criterion and many clustering methods are well suited to it.

In what follows, n will denote the number of nodes, where the LMP prices are computed, and c_{it} will be the LMP price in node $i = 1, \ldots, n$ at hour $t = 1, \ldots, T$. To allow different nodes to have different significance at different time, the weighting factors $w_{i,t}$ can be introduced. One of the standard approaches to weighting is to take the weights equal to the traded volumes $w_{i,t} = A_{i,t}$. Alternatively, one can assign a set of constant weights equal to the installed capacity in the nodes or equal to the annual average traded volume.

We will say that a node i is assigned to hub j, if the market participant located at node i uses hub j (and only this hub) for hedging. The set of nodes, assigned to a hub j will be called the *target region* of the hub j. In practice, a market participant may hold a set of nodes of the grid and trade the electricity in all of these nodes with certain proportion of the traded volumes in the nodes.

3.1 Construction of a Given Number of Hubs

The problem formulation considered here is based on the assumption that a given number of hubs m, $m < n$ is sufficiently small to ensure sufficient liquidity of hubs. The goal is to minimize the total (weighted) deviation D of the hub prices from the nodal prices in the target regions of the hubs. In effect, this means that for simplicity we assume that each node belongs to one market participant and each participant holds one node. The mathematical formulation of this problem is as follows:

$$\text{Min} \quad D = \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_i^j \sum_{t=1}^{T} (c_{it} - c_t^j)^2 w_{it} \tag{13}$$

s.t.

$$c_t^j = \sum_{i=1}^{n} \alpha_i^j c_{it}, \ j = 1, \ldots, m, \ t = 1, \ldots, T, \tag{14}$$

$$\sum_{i=1}^{n} \alpha_i^j = 1, \ j = 1, \ldots, m, \tag{15}$$

$$\sum_{j=1}^{m} \delta_i^j = 1, \ i = 1, \ldots, n, \tag{16}$$

$$\delta_i^j \in \{0, 1\}, \quad \alpha_i^j \geq 0, \ i = 1, \ldots, n, \ j = 1, \ldots, m. \tag{17}$$

Here the variables δ_i^j, $i = 1, \ldots, n$, $j = 1, \ldots, m$ define which nodes are assigned to each hub, the variables α_i^j, $i = 1, \ldots, n$ define the set of weights within the hub j and the variables c_t^j, $t = 1, \ldots, T$ give the price of hub j, $j = 1, \ldots, m$ at each hour t. Equation (14) gives the hub price calculation, while equation (16) ensures that each node is assigned to exactly one of the hubs. The constraint (15) serves for normalization of the hub price. Although the model would be meaningful without this constraint, in certain conditions it plays the role of a cut, as it will be seen in the proof of Proposition 1 below.

The Boolean variables δ_i^j may be substituted by real-valued variables ranging from 0 to 1. Although this relaxation of problem (13) – (17) allows each node to be assigned to several hubs simultaneously, it is easy to see that the relaxed formulation always has an optimal solution with Boolean values of all δ_i^j. (For each i one can assign $\delta_i^j = 1$ for a single hub j which minimizes $\sum_{t=1}^{T} (c_{it} - c_t^j)^2 w_{it}$.) This problem in the relaxed version was originally formulated by W. Hogan [15] for the case of two hubs and extended to optional number of hubs in [10], Appendix A.

As it is noted in [15], the relaxed formulation belongs to the class of non-convex optimization problems, thus it is impossible to apply directly the efficient optimization techniques developed in convex optimization. However, once the set of all variables δ_i^j is fixed, the remaining variables may be found by solving a convex optimization problem; sometimes they may be assigned

explicitly as we will see in the proof of Proposition 1 below. Alternatively, if the set of all variables α_i^j is given, the complementary assignment of the variables δ_i^j is straightforward. These properties may be exploited in the non-convex optimization algorithms [18], if they are tailored for this problem.

Note that in a feasible solution one or several hubs may have empty target regions, i.e. for these hubs j holds

$$\sum_{i=1}^{n} \delta_i^j = 0.$$

We will call such assignments *degenerate*. Note that it is possible to eliminate the empty target regions, not increasing the objective function value. This can be done iteratively by finding a node k with the maximal value of

$$\sum_{j=1}^{m} \delta_k^j \sum_{t=1}^{T} (c_{kt} - c_t^j)^2 w_{kt}$$

and assigning it to a hub with an empty target region. The new assignment, coupled with the available set of real-valued variables α_i^j and c_t^j, gives a feasible solution and does not increase the previously found value of objective function (in fact this holds for any choice of k).

The following proposition is aimed at finding the best-possible set of real-valued variables, complementing a non-degenerate assignment of nodes.

Proposition 1. *Suppose, $w_{it} = w_i$ does not depend on t for all nodes i and a feasible non-degenerate assignment $\{\delta_i^j\}$ is given. Then the optimal price in hub j, $1 \leq j \leq m$ is calculated as the weighted average of prices in the assigned nodes:*

$$c_t^j = \sum_{i:\ \delta_i^j = 1} w_i c_{it} \Big/ \sum_{i:\ \delta_i^j = 1} w_i. \tag{18}$$

Proof. Denote by F_{jt} the deviation of nodal prices in hub j at hour t.

$$F_{jt} = \sum_{i=1}^{n} \delta_i^j (c_{it} - c_t^j)^2 w_i.$$

To find hub price c_t^j minimizing F_{jt} we differentiate it over c_t^j:

$$\frac{\partial F_{jt}}{\partial c_t^j} = -\sum_{i=1}^{n} \delta_i^j 2(c_{it} - c_t^j) w_i = 0.$$

Solving this equation we obtain

$$c_t^j = \sum_{i=1}^{n} \delta_i^j w_i c_{it} \Big/ \sum_{i=1}^{n} \delta_i^j w_i.$$

Denote $\alpha_i^j = w_i / \sum_{k=1}^{n} \delta_k^j w_k$ if $\delta_i^j = 1$ and $\alpha_i^j = 0$ otherwise. It is easy to check that $\sum_{i=1}^{n} \alpha_i^j = 1$, so the obtained solution is feasible and hence it is optimal.   Q.E.D.

In conditions of Proposition 1, hub j, $1 \le j \le m$ is completely defined by its target region

$$H_j = \{i : \delta_i^j = 1, \ i = 1, \ldots, n\},$$

since the coefficients α_i^j are given by

$$\alpha_i^j = w_i \Big/ \sum_{k: \ \delta_k^j = 1} w_k.$$

In the special case where the weights w_i are all identical, the problem turns into the classical minimum sum-of-squares clustering problem: find a partition of a given finite set of vectors in Euclidean space into several disjoint sets (clusters), minimizing sum of squared distances from each element to the centroid of its cluster. Here centroid means the simple average of vectors in a cluster. Currently the complexity status of this problem is open, in spite of a number of attempts to prove that this problem is NP-hard (see the survey [1]). This problem has been deeply studied during the last 50 years and a number of exact and heuristic approaches to its solution have been developed (see the survey in [11]). Some of them will be discussed in Section 4.

In the case of identical weights w_i, the optimal hub price (18) equals to the simple average of the nodal prices, which makes it appropriate to impose a lower bound $n_{\min}$, $n_{\min} \le n/m$ on the number of nodes in each hub:

$$n_{\min} \le \sum_{i=1}^{n} \delta_i^j, \quad j = 1, \ldots, m. \tag{19}$$

This modification of the problem is not studied as much as the minimum sum-of-squares clustering problem and its solution may require some modification of the well-known clustering methods or application of general-purpose optimization tools.

Let us consider what modification of the objective function (13) is required in order to minimize the total observed variance (6), assuming that participants use the hedge ratio approach described above. With simplifying assumption that the prices are stationary distributed, the estimated variance $\hat{\sigma}_i^2$ of price c_{it} in node i, as well as the estimated correlation $\hat{\rho}_{ij}$ between the nodal price c_{it} and the hub price c_t^j, can be expressed on the basis of the historical data. Then equation (11) leads to the following criterion:

$$\text{Min} \ \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_i^j \hat{\sigma}_i^2 \left(1 - \hat{\rho}_{ij}^2\right), \tag{20}$$

where

$$\hat{\sigma}_i^2 = \sum_{t=1}^{T} \left(c_{it} - \frac{1}{T} \sum_{t=1}^{T} c_{it} \right)^2 \Big/ (T-1),$$

$$\hat{\rho}_{ij}^2 = \frac{\left(T \sum_{t=1}^{T} c_{it} c_t^j - \sum_{t=1}^{T} c_{it} \sum_{t=1}^{T} c_t^j \right)^2}{\left(T \sum_{t=1}^{T} c_{it}^2 - \left(\sum_{t=1}^{T} c_{it} \right)^2 \right) \left(T \sum_{t=1}^{T} (c_t^j)^2 - \left(\sum_{t=1}^{T} c_t^j \right)^2 \right)}$$

for $i = 1, \ldots, n, \ j = 1, \ldots, m$.

3.2 Single Hub Selection

The purpose for formulation of the Single Hub Selection Problem in this section is to refine a set of m preliminary hubs by selecting a refined hub within each of them. The set of preliminary hubs may be a result of selecting a given number of hubs, or it may describe an existing set of hubs or zones in the electricity market.

   We can assume without loss of generality that the set of nodes of the preliminary hub is $\{1, \ldots, N\}$, where $N \leq n$. In what follows, talking about the Single Hub Selection Problem we will use the term "hub" only for the refined hub, while the preliminary hub will be referred to as *a given set of nodes* or *cluster* .

   We will assume that the hub is chosen with respect to the locational energy prices of the market participants situated in the target region of the hub. Let p_{rt} denote the energy price of participant r, $r = 1, \ldots, R$ at hour t. A particular definition of the price of participant does not matter. In case a participant r has the injection/withdrawal of energy within a single node i of the grid, the energy price of this participant usually equals c_{it}. If the injection/withdrawal of a participant is spread over a number of nodes, then the price p_{rt} may be calculated as a weighted average of the nodal prices according to some market rules.

   Suppose the hub price is always computed as an average price over all included nodes, and require that the hub contains at least $n_{\min}$ nodes. Then the Single Hub Selection Problem consists in minimizing the sum of squared differences of the prices of participants from the hub price with respect to a given set of weights of market participants W_{rt}, $r = 1, \ldots, R$:

$$\text{Min } f = \sum_{t=1}^{T} \sum_{r=1}^{R} (c_t - p_{rt})^2 W_{rt} \tag{21}$$

s.t.

$$c_t = \frac{1}{L} \sum_{i=1}^{N} x_i c_{it}, \ t = 1, \ldots, T, \tag{22}$$

$$\sum_{i=1}^{N} x_i = L, \tag{23}$$

$$L \geq n_{\min}, \tag{24}$$

$$x_i \in \{0, 1\}, \ i = 1, \ldots, N, \ c_t \geq 0, \ t = 1, \ldots, T. \tag{25}$$

Here the binary variables x_i turn into 1 whenever node i is included into the hub. The variables c_t define the hub price at time t, $t = 1, \ldots, T$. The complexity status if this problem in the special case, when each participant is located in its own node, is established by the following proposition.

Proposition 2. *The Single Hub Selection Problem (21)–(25) is NP-hard even when $R = N$, $p_{it} = c_{it}$ for all $i = 1, \ldots, N$, $t = 1, \ldots, T$ and $T = 2$.*

The proof of Proposition 2, provided in the Appendix, is based on a transformation from an NP-complete Partition problem.

4 Heuristics for Construction of Given Number of Hubs

In this section, we discuss two well-known clustering heuristics in the context of the Hubs Construction Problem. We assume that all nodes are given constant weights w_i. According to Proposition 1, it is sufficient to partition the set of nodes $\{1, \ldots, n\}$ into m clusters $H_1, \ldots, H_m$, minimizing the total weighted squared error:

$$\text{Min} \quad S = \sum_{j=1}^{m} \sum_{i \in H_j} w_i \sum_{t=1}^{T} (c_{it} - c_t^j)^2,$$

where the hub price c_t^j is calculated as the weighted average of nodal prices in cluster H_j:

$$c_t^j = \frac{\sum_{i \in H_j} c_{it} w_i}{\sum_{i \in H_j} w_i}. \tag{26}$$

4.1 The H-Means Method

R. Howard [19] is considered to be the first one who outlined the clustering method H-means. Here we view this algorithm in adaptation to the Hubs Construction Problem with a given number of hubs. Starting with an initial set of points c_j, $j = 1, \ldots, m$ in T-dimensional Euclidean space, H-means algorithm iterates the following three steps:

1. For each node i, $i = 1, \ldots, n$, find the closest c_j, $j \in \{1, \ldots, m\}$ w.r.t. Euclidean distance and place the node i into the cluster H_j.
2. Let m' be the number of non-empty clusters and reorder the clusters so that $H_1, \ldots, H_{m'} \neq \emptyset$.

3. Recalculate c_j for all $j = 1, \ldots, m'$ according to (26) with the new partition $H_1, \ldots, H_{m'}$.

The algorithm terminates when the set of clusters does not change any more. If during the run of the algorithm the number of non-empty clusters m' falls below m, the assignment of nodes becomes degenerate.

In the case when all weights w_i are identical, it is well known [9] that Step 1 gives the optimal partition for the given centroids, and Step 3 gives the optimal centroids location for the given partition. In view of Proposition 1, it is easy to see that the same holds if w_i are not identical.

At each iteration, the value of objective function can not increase, so the algorithm will eventually reach some value of objective function it can not further improve. This will take only a finite number of iterations because there is only a finite number of partitions of a finite data set. The computational cost of each iteration is equal to $O(mnT)$.

The output of H-means depends on the initial set of centroids $c_1, \ldots, c_m$ and it is not necessarily a global optimum of problem (13) – (17). The initial values $c_1, \ldots, c_m$ may be provided by an expert in the form of cluster seeds (a set of clusters, each consisting of a single node) to direct the heuristic to some "reasonable" structure of hubs. Alternatively, one can run the H-means algorithm a number of times with different randomly chosen cluster seeds and choose the best output over all runs.

The computational study in [12] indicates an advantage of the following simple modification of H-means. The difference of the modification, named H-means+, from the original method consists in checking for degeneracy of the assignment found. The algorithm stops if the assignment is not changing any more and it is non-degenerate. If it is degenerate, the number of non-empty clusters is raised up to m as it was described in Section 3.1 and the iterations continue.

Minimization of Euclidean distance may be substituted by other criteria mentioned in Section 3, e.g. objective (20). The H-means algorithm is sufficiently flexible and it may be adjusted to use such criteria as well (on applicability of H-means see [9, 32] and references therein).

4.2 The K-Means Heuristic

R. Jancey [21] and J. MacQueen [23] proposed the K-means heuristic which is similar to the H-means but fits better into the standard local search scheme. Here we use the terminology of Hubs Construction Problem, presenting a slightly generalized version of H-means, which takes the nodal weights w_i into account.

The K-means starts from an initial partition $H_1, \ldots, H_m$ of nonempty hubs and iteratively moves a node from one hub to another to decrease the value of objective function D. To choose a node to be moved, all possible reassignments are considered and the one with largest decrement of the objective

function value is chosen. The iterations are performed until either no nodes can be moved, or the value of D decreases unsubstantially.

Without loss of generality, suppose that node k is moved from H_1 to H_2. Hubs $H_3, \ldots, H_m$ are unaffected by the move, H_1 transforms into $\tilde{H}_1 = H_1 \setminus \{k\}$, and H_2 becomes $\tilde{H}_2 = H_2 \cup \{k\}$. Denote $c_i = (c_{i1}, \ldots, c_{iT})$, $c^j = (c_1^j, \ldots, c_T^j)$ and

$$\|c_i - c^j\|^2 = \sum_{t=1}^{T} (c_{it} - c_t^j)^2$$

and find the difference between total weighted squared errors analogously to the computations in [31]:

$$
\begin{aligned}
\tilde{D} - D &= \sum_{i \in \tilde{H}_1} w_i \|c_i - \tilde{c}^1\|^2 + \sum_{i \in \tilde{H}_2} w_i \|c_i - \tilde{c}^2\|^2 \\
&\quad - \sum_{i \in H_1} w_i \|c_i - c^1\|^2 - \sum_{i \in H_2} w_i \|c_i - c^2\|^2 \\
&= \sum_{i \in H_1} w_i \|c_i - \tilde{c}^1\|^2 + \sum_{i \in H_2} w_i \|c_i - \tilde{c}^2\|^2 \\
&\quad - \sum_{i \in H_1} w_i \|c_i - c^1\|^2 - \sum_{i \in H_2} w_i \|c_i - c^2\|^2 \\
&\quad - w_k \|c_k - \tilde{c}^1\|^2 + w_k \|c_k - \tilde{c}^2\|^2 \\
&= \sum_{i \in H_1} w_i \left(\|c_i - \tilde{c}^1\|^2 - \|c_i - c^1\|^2 \right) \\
&\quad + \sum_{i \in H_2} w_i \left(\|c_i - \tilde{c}^2\|^2 - \|c_i - c^2\|^2 \right) \\
&\quad - w_k \|c_k - \tilde{c}^1\|^2 + w_k \|c_k - \tilde{c}^2\|^2.
\end{aligned}
$$

Due to the equality

$$
\begin{aligned}
\|c_i - b\|^2 - \|c_i - a\|^2 &= \sum_{t=1}^{T} (a_t - b_t)(2c_{kt} - a_t - b_t) \\
&= 2 \sum_{t=1}^{T} (a_t - b_t)(c_{kt} - a_t) + \|b - a\|^2,
\end{aligned}
$$

one has

$$
\begin{aligned}
\sum_{i \in H_1} & w_i \left(\|c_i - \tilde{c}^1\|^2 - \|c_i - c^1\|^2 \right) \\
&= \sum_{i \in H_1} w_i \|\tilde{c}^1 - c^1\|^2 + 2 \sum_{i \in H_1} w_i \sum_{t=1}^{T} (c_t^1 - \tilde{c}_t^1)(c_{it} - c_t^1) \\
&= \|\tilde{c}^1 - c^1\|^2 \sum_{i \in H_1} w_i + 2 \sum_{t=1}^{T} (c_t^1 - \tilde{c}_t^1) \sum_{i \in H_1} w_i (c_{it} - c_t^1).
\end{aligned}
$$

By definition of the hub price,

$$\sum_{i \in H_1} w_i(c_{it} - c_t^1) = 0.$$

Similar calculations are used for other summands to obtain,

$$\tilde{D} - D = \|\tilde{c}^1 - c^1\|^2 \sum_{i \in H_1} w_i + \|\tilde{c}^2 - c^2\|^2 \sum_{i \in H_2} w_i - w_k\|c_k - \tilde{c}^1\|^2 + w_k\|c_k - \tilde{c}^2\|^2.$$

Therefore, one chooses the reassignment providing the minimal value of $\tilde{D} - D$ among all possible moves of each node.

The computational cost of one iteration is equal to $O(mnT)$, as well as in the H-means. This algorithm may be restarted a number of times from randomly chosen partitions.

The K-means and H-means clustering methods are probably the most widely used in practice. This is due to the simplicity and computational efficiency of these algorithms. The quality of their solutions, however, may be far from the optimal. This is demonstrated, e.g. in [12], where both of these algorithms were experimentally compared to the Variable Neighborhood Search. A significant improvement of output results of K-means and H-means is reported for a combined method, which firstly starts the H-means+, and the obtained solution is further optimized by the K-means. This finding is consistent with the fact that any solution, which is non-improvable for the K-means, is also non-improvable for the H-means, while the converse is not true [34].

We have considered only two well-known heuristics adapted to the Hubs Construction Problem. A number of other exact and heuristic approaches, such as the branch-and-bound algorithms, metaheuristics, hierarchical clustering heuristics and other methods (see e.g. [11, 12] and references there) can be also applied to the Hubs Construction Problem and its modifications. However, the main limiting factor, which may hinder the usage of some of these methods, is a large dimensionality of typical instances of the Hubs Construction Problem. A generalization of the known methods to the case of time-dependent nodal weights w_{it} constitutes another challenge for algorithmic research.

4.3 Experimental Evaluation of the K-Means on PJM Data

Performance of the K-means method described in Section 4.2 is tested here on the Real-Time Market hourly prices from PJM Interconnection, available at `http://www.pjm.com`. The input includes hourly data of 5 weeks, each week representing one of the months from January to May of year 2007, in total 840 records for each of $n = 7599$ nodes. Analogously to the existing 11 PJM hubs, $m = 11$ is chosen. All nodal weights are set to 1.

The K-means is programmed in C++ and tested on Pentium-IV, 3 GHz machine. A series of 30 independent runs is made with random initial solutions, each run taking from 40 min to 1 hour. The best outcome in terms of

the objective function (13) consists of 11 hubs with sizes ranging from 152 to 1898 nodes. It turns out that if the price of the new hubs were computed as the simple average of their nodal prices, then our largest hub would be the closest one to the existing Western hub. We denote this hub by LS-Hub1. The closest to LS-Hub1, among the existing hubs, is Dominion.

An 825-node hub, closest to PJM Eastern hub, we denote by LS-Hub2. It also tightly approximates the existing New Jersey hub. The hubs AEP Gen, AEP Dayton, Chicago Gen, Chicago, Northern Illinois and Ohio are approximated by other hubs found by the K-means. The Western Interface has no equivalent in the set of our hubs. At the same time the set of our hubs contains a high-price hub of 313 nodes with no equivalent among the existing hubs.
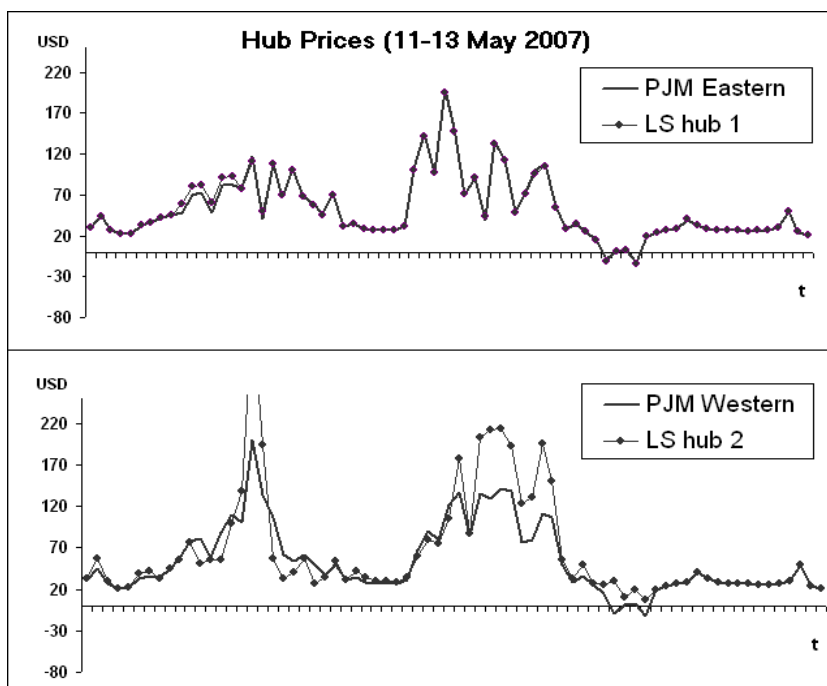

Fig. 2. Energy prices in PJM Western Hub, PJM Eastern Hub, LS-Hub1 and LS-Hub2 on May 11–13, 2007

Figure 2 demonstrates the behavior of prices in Western, Eastern, LS-Hub1 and LS-Hub2 during tree days from Friday, May 11 to Sunday, May 13, 2007. As it can be seen from the figure, the prices of PJM Eastern and LS-Hub2 are very close. However, PJM Western hub does not have the price peaks as high as the peaks of its counterpart LS-Hub1. The graph of Dominion exactly overlaps with LS-Hub1, so we not display it on the figure.

Absence of exact counterpart for the PJM Western hub in the output of the K-means heuristic may be due to the difference of clustering methods and their settings, different historical data and subsequent optimization and modifications of PJM Western hub, which followed the initial clustering stage. A higher volatility of our largest hub LS-Hub1, compared to PJM Western, may lead to lower liquidity of LS-Hub1 because the price of the latter may be more difficult to predict. At the same time, the companies located in BGE, Dominion and PEPCO, experiencing high on-peak prices could use the LS-Hub1 more actively for hedging their contracts, which is a positive factor for the hub liquidity.

This computational experiment illustrates that even a simple K-means clustering can produce a meaningful initial grouping of nodes. The running time of the K-means is not exceptionally high even for a system of about 7000–8000 nodes and the same approach could be applied to the data representing all 12 months of the year.

5 Solving the Single Hub Selection Problem

The non-linear discrete optimization problem (21)–(25) can be transformed into a family of Boolean linear programming problems, each one with a different value of parameter L, $L = n_{\min}, \dots, N$. Let us consider a term $(c_t - p_{rt})^2$ from (21) for any pair of r and t:

$$(c_t - p_{rt})^2 = \left(\frac{1}{L} \sum_{i=1}^{N} x_i c_{it} - p_{rt} \right)^2 =$$

$$\frac{1}{L^2} \sum_{i=1}^{N} (c_{it}^2 - 2L c_{it} p_{rt}) x_i + \frac{2}{L^2} \sum_{k=1}^{N} \sum_{l=1}^{k-1} c_{kt} c_{lt} x_k x_l + p_{rt}^2. \tag{27}$$

In view of this expression, one can remove the non-linearity from the objective function (21) by introducing new variables y_{kl}, $k = 1, \dots, N$, $l = 1, \dots, k-1$ so that

$$y_{kl} = x_k x_l, \ k = 1, \dots, N, \ l = 1, \dots, k-1. \tag{28}$$

The set of equalities (28) may also be substituted by a system of linear constraints:

$$y_{kl} \leq x_k, \quad y_{kl} \leq x_l, \quad k = 1, \dots, N, \ l = 1, \dots, k-1, \tag{29}$$

$$y_{kl} \geq x_k + x_l - 1, \quad k = 1, \dots, N, \ l = 1, \dots, k-1. \tag{30}$$

Therefore, the Single Hub Selection Problem with given value L becomes a mixed integer linear programming (MIP) problem. In view of (27), it is easy to notice that constraints (29) are always satisfied in the optimum, even if

they were not included into problem formulation. Now we can conclude that problem (21)–(25) reduces to solving the following family of MIP problems:

$$\text{Min} \ \ C_0 + \sum_{i=1}^{N} C_i x_i + \sum_{k=1}^{N} \sum_{l=1}^{k-1} B_{kl} y_{kl} \tag{31}$$

s.t.

$$y_{kl} \geq x_k + x_l - 1, \ \ k = 1, \ldots, N, \ l = 1, \ldots, k-1, \tag{32}$$

$$\sum_{i=1}^{N} x_i = L, \tag{33}$$

$$x_i \in \{0, 1\}, \ \ i = 1, \ldots, N, \tag{34}$$

$$y_{k,l} \geq 0, \ \ k = 1, \ldots, N, \ l = 1, \ldots, k-1, \tag{35}$$

where $L \in \{n_{\min}, \ldots, N\}$ and the coefficients of objective function are:

$$C_0 = \sum_{r=1}^{R} \sum_{t=1}^{T} p_{rt}^2 W_{rt},$$

$$C_i = \frac{1}{L^2} \sum_{r=1}^{R} \sum_{t=1}^{T} (c_{it}^2 - 2L c_{it} p_{rt}) W_{rt},$$

$$B_{kl} = \frac{2}{L^2} \sum_{r=1}^{R} \sum_{t=1}^{T} c_{kt} c_{lt} W_{rt}.$$

Indeed, if one selects the value L which yields the optimum with minimal objective function among all problems (31)–(35) of the family, it will be the optimal solution to problem (21)–(25) as well.

An important property of this MIP formulation is that now the time dimension T and the total number of market participants R do not influence the dimensionality of the model because these parameters are excluded from consideration at the stage of computing the coefficients C_i and B_{kl}. This fact becomes important, e.g. when the historical data consist of the nodal prices of all hours of the previous year. Taking into account that when the number of variables is bounded, the MIP problems fall into the class of polynomially solvable problems [22], we conclude that the Single Hub Selection Problem is also polynomially solvable, if the number of nodes is bounded above by a constant.

5.1 Genetic Algorithm

Genetic algorithm (GA) originally proposed by J. Holland [17] is a random search method that models a process of evolving a population of individuals. Each individual corresponds to some solution of the problem (feasible

or maybe infeasible) and it is characterized by the *fitness*, reflecting the goal function value and satisfaction of problem constraints. The higher is the fitness value, the more chances are given for the individual to be selected as a parent. New individuals are built by means of *crossover* and *mutation* procedures. The crossover procedure $Cross$ produces the offspring from two parent individuals by combining and exchanging their elements. The mutation procedure Mut adds small random changes to an individual. The size of population K is kept constant throughout the run of a GA. A detailed description of the GAs and their properties may be found e.g. in [28].

For solving the Single Hub Selection Problem we use the *binary representation* of solutions in the GA, i.e. an individual in our case is a string g which coincides with the Boolean N-dimensional vector x. The fitness of individual is inversely proportional to the objective function value. Parent genotypes are selected by *s-tournament* selection operator: choose s individuals from the population at random and return the best of them (by default in this section "random" means random with uniform distribution). This selection operator is used to choose each of the two parents independently. New individuals are produced by the *2-point* crossover operator, which chooses randomly two breakpoints in parent genotypes and exchanges all bits in the middle part. The standard mutation inverses each bit independently with a fixed probability p_m. If after crossover and mutation the obtained genotype contains less than $n_{\min}$ ones then a repair procedure is applied. This procedure simply adds more ones to the child individuals at random positions. The overall scheme of the GA used here is as follows:

Genetic Algorithm
1. Generate K random genotypes and add into the initial population.
2. While the termination condition is not met, do
    2.1. Choose the parent genotypes g_u, g_v by s-tournament selection.
    2.2. Produce g, h from g_u and g_v
        using 2-point crossover with probability p_c,
        otherwise assign $g = g_u$, $h = g_v$.
    2.3. Mutate each gene of g and h with probability p_m.
    2.4. Apply repair procedure to g and h.
    2.5. Choose two individuals of least fitness in the current population
        and substitute them by g and h, if they have greater fitness.
3. Return the best found solution as a result.

5.2 Experiments with the GA and CPLEX MIP-Solver

The genetic algorithm is tested here on the hourly electricity prices over 365 days from the day-ahead two-sector electricity market of the European zone of Russia collected in years 2004-2005. First of all, the H-means heuristic is applied to form a set of clusters (preliminary hubs), using identical

weights $w_i = 1$, $i = 1, \ldots, n$. The GA is applied to form one hub in each cluster. The problem characteristics and the results are given in Table 1. Here the larger instances P3, P4, and P5 correspond to clusters located in Urals-Volga, Urals-Tyumen, and Center regions accordingly. The smaller ones, P1 and P2, are constructed as random subsets from the cluster of P3. The instances P6, P7, and P8 are based on the same clusters as P3, P4 and P5, but considered over half-year (4343 hours) time horizon. The required minimal number of nodes $n_{\min}$ for the GA is set to $[N/2]$ (here the brackets $[\cdot]$ denote rounding to the nearest integer). In our experiments, we set the following control parameters: $s = 20$, $p_c = 1/2$, $K = 200$ and $p_m = 1/N$. The actual number of nodes L in the computed hubs turns out to be equal or close to $n_{\min}$. This value of L is used in problem formulation (31)–(35), which is also solved by CPLEX 11.0. The amount of CPU time at Celeron 2.8 GHz is indicated in the table as well. After this time both algorithms are terminated.

The results show a clear advantage of the GA in terms of the running time and the solution quality. The lower bounds obtained by CPLEX in the given amount of time are negative in most of the cases and they cannot be of practical use. For problem P8 CPLEX fails because of memory limitation.

Table 1. Comparison of the GA and CPLEX 11.0

Problem	N	R	T	CPLEX CPU time	CPLEX best sol	CPLEX lo. bound	GA CPU time	GA best sol
P1	15	14	500	15 sec.	26.97	26.97	<1 sec.	26.97
P2	25	14	500	5 min.	22.97	-8641	<1 sec.	22.95
P3	82	14	500	5 min.	24.53	-69230	10 sec.	23.24
P4	118	15	500	5 min.	23.8	-62137	30 sec	23.53
P5	336	69	500	10 min.	16.19	-72755	2 min	14.46
P6	82	14	4343	30 min.	196.55	-519978	5 min.	146.28
P7	118	15	4343	60 min.	389.80	-545399	5 min.	364.25
P8	336	69	4343	-	-	-	10 min.	208.00

Evaluation and comparison of the obtained hubs.

In the clusters corresponding to P6, P7, and P8 the following different hubs are constructed and compared:

1. A hub constructed by the GA minimizing quadratic objective (21).
2. A hub constructed by the GA minimizing linear objective

$$\sum_{t=1}^{T} \sum_{r=1}^{R} |c_t - p_{rt}| W_{rt}. \tag{36}$$

3. A hub constructed by the GA *maximizing* linear objective (36). This hub gives a worst case in a linear model (for the sake of comparison only).

4. A hub containing all nodes of a cluster.
5. A hub containing a randomly chosen subset of nodes of a cluster (every node is included independently of the other nodes with probability 0.5).

The set of conditions (23)–(25) is never changed. The comparison is illustrated by Fig. 3. Here each hub is represented by a point on a plane where X and Y axes correspond to the values of linear and quadratic objective functions. The results show that the optimized hubs are far from the worst case hub in
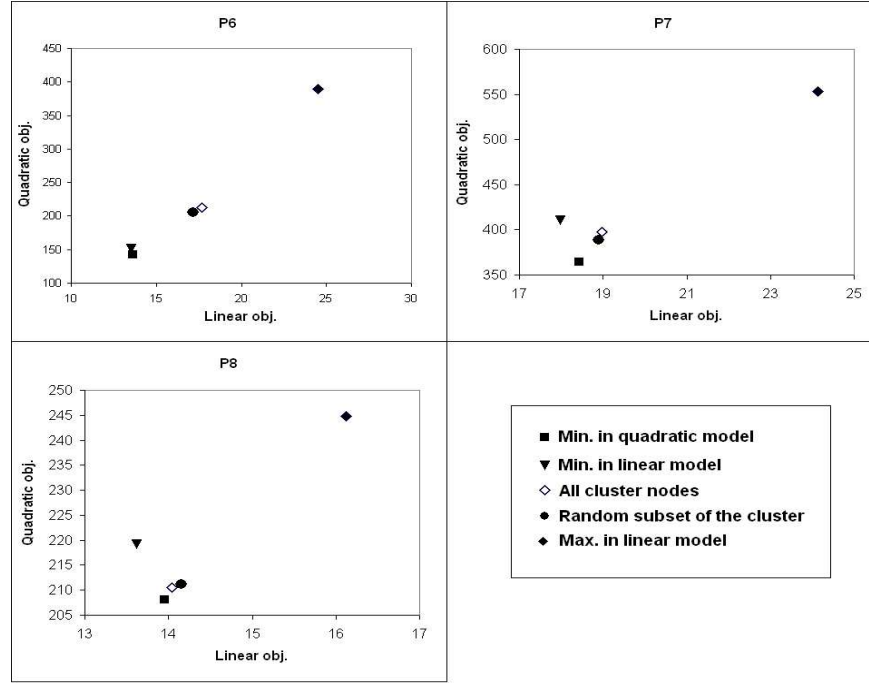

Fig. 3. Evaluation and comparison of different solutions to Single Hub Selection Problem

terms of both criteria and not so much distant from each other. In the cases of P7 and P8, the hub optimized w.r.t. linear objective (36) has even greater value of quadratic objective (21) than the hub consisting of all nodes of the cluster. This indicates that the choice of optimization criterion is important and it should be adjusted to the interests of participants.

Figure 3 implies that some proper subsets of the cluster can constitute better hubs than the set of all nodes. The question about the trade-off between the number of nodes in hub and the attainable objective value is addressed in the next experiment, where we ran the GA minimizing quadratic objective separately for different values of $n_{\min}$. The results are shown on Fig. 4. One

can see that a fast growth begins near the point $n_{\min} = 40$, so this setting is a plausible candidate to form a hub. In practice, a graph of such type can provide useful information for the decision-maker, who needs to choose a sufficiently large value of $n_{\min}$ without significant compromise in distance between the hub price and the locational prices of participants.
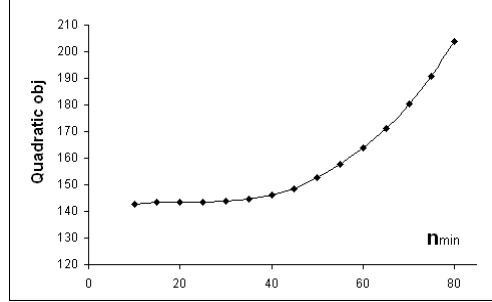

Fig. 4. Trade-off between the number of nodes in hub and the objective value

6 Conclusion

We have considered the Hubs Construction Problem from the optimizational prospective. Our analysis of this problem indicated that even though its connections with clustering problems allow to apply the well-known clustering methods, some important constraints and criteria do not necessarily fit into the clustering framework. In view of large dimensionality of typical instances of the Hubs Construction Problem, further development of the specialized optimization algorithms is important to support decision making.

A set of heuristic algorithms, we have considered, allows to find practically useful solutions. Even if the clustering algorithm does not yield an acceptable solution for most of the market participants, each of the obtained hubs can be further refined by solving a Single Hub Selection Problem. A genetic heuristic is shown to be suitable for finding approximate solutions to this problem with different criteria.

An important direction for further research is evaluation of the Principal Component Analysis and its comparison to the optimization-based methods described in this chapter. In some situations the hubs need to be defined even before the new electricity market opens. The statistical data on LMP prices is absent in such cases and one has to construct the hubs using some kind of market simulation and/or the statistical data describing the transmission of the electric power in the grid. This presents another challenge for research. The issues of hubs construction in view of negative influence of the market power require a careful consideration as well.

Acknowledgement

Appendix

Proof of Proposition 2. The proof is by reduction from the following NP-complete version of Partition problem (see Appendix 3.2 in [8]): Given M integers $a_1, \ldots, a_M$, recognize the existence of such subset $I \subseteq \{1, \ldots, M\}$ of cardinality $M/2$ that $\sum_{i \in I} a_i = \frac{1}{2} \sum_{i=1}^{M} a_i$.

Given a set of integers $a_1, \ldots, a_M$, we construct an instance of the Single Hub Selection Problem with $T = 2$, $N = M + 2$, and $n_{\min} = M/2 + 1$.

We assign the data of hour 1 in such a way that any optimal solution H^* to the Single Hub Selection Problem (i) will consist of $M/2 + 1$ nodes, (ii) it will contain the node number $M + 1$ and (iii) it will not contain the node number $M + 2$. To this end we put $W_{i,1} = 0$, $i = 1, \ldots, M$; $W_{M+1,1} = 1$, $W_{M+2,1} = 0$; $c_{i,1} = K$, $i = 1, \ldots, M$; $c_{M+1,1} = 0$, $c_{M+2,1} = 2K$, where the parameter K is sufficiently large (its value will be chosen later). Note that with these assumptions the price of a hub H at hour 1 will be

$$c_1 = \begin{cases} K, & \text{if } M+1 \notin H, \ M+2 \notin H; \\ K(1 - 1/|H|), & \text{if } M+1 \in H, \ M+2 \notin H; \\ K(1 + 1/|H|), & \text{if } M+1 \notin H, \ M+2 \in H; \\ K, & \text{if } M+1 \in H, \ M+2 \in H. \end{cases} \tag{37}$$

At the same time, the only non-zero term, associated with hour 1 in sum (21) equals $(c_1)^2$. Thus, if K is sufficiently large and the input data for hour 2 does not depend on K, then the optimal hub will always meet conditions (ii), (iii) and involve the minimal admissible number of nodes, i.e. condition (i) holds as well.

Now we proceed to the input data of hour 2, which will ensure equivalence of (21)–(25) to the given Partition problem, assuming that conditions (i)–(iii) are satisfied. Let $c_{i,2} = a_i(M/2 + 1)$, $i = 1, \ldots, M$; $c_{M+1,2} = 0$, $c_{M+2,2} = \frac{1}{2} \sum_{i=1}^{M} a_i$. Then in hour 2 the hub price for an optimal hub H^* will be $c_2 = \sum_{i \in H^*} a_i$. Finally, assign $W_{i,2} = 0$, $i = 1, \ldots, M + 1$ and $W_{M+2,2} = 1$. It is clear that to ensure (i)–(iii) it is sufficient to assign $K = M \sum_{i=1}^{M} a_i$.

Note that the only non-zero summand in (21) at hour $t = 2$ is $(c_2 - c_{M+2,2})^2$, which attains the minimum equal to 0 if and only if $\sum_{i \leq M, \ i \in H^*} a_i = \frac{1}{2} \sum_{i=1}^{M} a_i$. So, the optimal value of criterion (21) is equal to $K(1 - 2/(M+2))$

if and only if there exists the set I required in the Partition problem. Thus, we have reduced an NP-complete problem to the Single Hub Selection Problem, and the reduction can be computed in polynomial time. Q.E.D.

References

1. Aloise D, Hansen P (2007) On the complexity of minimum sum-of-squares clustering. Les Cahiers du GERAD, G-2007-50, GERAD, Montréal
2. Braziel E R (1998) Trading hubs: Where power is traded and why. PMA OnLine Magazine 12
3. Campbell J F, Ernst A Krishnamoorthy M (2002) Hub location problems. In: Hamacher H, Drezner Z (eds) Location analysis: Theory and applications. Springer, Berlin
4. Cai X (2005) An evaluation of investment incentives for maintaining system adequacy in deregulated markets for electricity. PhD dissertation, Cornell University
5. Davidson M P, Dogadushkina U V, Kreines E M, Novikova N M, Udaltsov U A, Shirjaeva L V (2004) Mathematical model of competitive wholesale electricity market in Russia. Izvestija Akademii Nauk. Teorija i Sistemi Upravlenija 3: 72-83 (in Russian)
6. Derivatives and risk management in the petroleum, natural gas, and electricity industries (2002) In: Derivatives and risk management in energy industries. Report SR-SMG/2002-01, EIA, Washington, DC
7. Falbo P, Stefani S (2004) Use and misuse of price indexes in the electricity markets. Working Papers Dipartamento Metodi Quantitativi Universitá di Milano Bicocca, Milan
8. Garey M R, Johnson D S (1979) Computers and intractability. A guide to the theory of NP-completeness. W.H. Freeman and Company, San Francisco
9. Gersho A, Gray R (1992) Vector quantization and signal compression. Kluwer Academic Publishers, Norwell, MA
10. Hartshorn A, Chang S (2003) MWISO hubs development. LECG, LLC, Cambridge, MA
11. Hansen P, Jaumard B (1997) Cluster analysis and mathematical programming. Mathematical Programming 79: 191–215
12. Hansen P, Mladenović N (2001) J-means: A new local search heuristic for minimum sum-of-squares clustering. Pattern Recognition 34: 405–413
13. Harvey S M, Hartshorn A P, Bertagnolli D and Kowalski R (1999) Derivation of a trading hub in the New England Market, NEPOOL Joint CMS/MSS Group
14. Hogan W W (1992) Contract networks for electric power transmission. J Regulatory Econom 4 (3): 211–242
15. Hogan W W (1998) Pivate communication
16. Hogan W W (2002) Financial transmission right formulations. Working Paper, Belfer Center for Business and Government, Cambridge, MA
17. Holland J H (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor
18. Horst R, Pardalos P M and Thoai N V (1995) Introduction to global optimization. Nonconvex optimization and its applications 3, Kluwer Academic Publishers, Dordrecht

19. Howard R (1966) Classifying a population into homogeneous groups. In: Lawrence J.R. (ed.) Operational research in the social sciences. Tavistock Publ. London
20. Isemonger A G (2005) CAISO trading hub white paper. August 16-18 2005 Meeting. California ISO
21. Jancey R C (1966) Multidimensional group analysis. Australian J Botany 14: 127–130
22. Lenstra H W Jr (1983) Integer programming with a fixed number of variables. Mathematics of Operations Research 8: 538–548
23. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proc. of the 5-th Berkeley Symposium on Mathematical Statistics and Probability 2: 281–297
24. NEPOOL energy market hub white paper and proposal (2003) Hub Analysis Working Group NEPOOL Markets Committee. October 2003
25. Ott A (2006) Private communication
26. Philpott A, Pritchard G (2004) Financial transmission rights in convex pool markets. Operations Research Letters 32: 109-113
27. Popova J (2004) Spatial pattern in modeling electricity prices: evidence from the PJM market. In: Proc. of 24th USAEE and IAEE North American Conference, July 8-10, Washington, DC
28. Reeves C, Rowe J (2003) Genetic algorithms – principles and perspectives. A guide to GA theory. Kluwer Academic Publishers, Boston
29. Shawky H A, Marathe A, Barrett C L (2003) A First look at the empirical relation between spot and futures electricity prices in the United States. The Journal of Futures Markets 23: 931–955
30. Schweppe F C, Caramanis M C, Tabors R D and Bohn R E (1988) Spot pricing of electricity. Kluwer Academic Publishers, Norwell, MA
31. Späth H (1980) Cluster analysis algorithms for data reduction and classification of objects. Chichester, Ellis Horwood
32. Selim S Z, Ismail M A (1984) K-means type of algorithms: A generalized convergence theorem and characterization of local optimality. IEEE Transactions on Pattern Analysis and Machine Intelligence 6(1):81–87
33. Summary standard market design notice of proposed rulemaking (2002) FERC Docket No. RM01-12-000
34. Zhang B, Kleyner G, Hsu M (1999) A local search approach to K-clustering. HP Labs Technical Report HPL-1999-119
35. Zwick W R, Velicer W F (1986) Comparison of five rules for determining the number of components to retain. Psychological Bulletin 99: 432–442