

# A Holistic Approach to Forecasting Wholesale Energy Market Prices

Ana Radovanovic, Tommaso Nesti, and Bokan Chen.

**Abstract**—Electricity market price predictions enable energy market participants to shape their consumption or supply while meeting their economic and environmental objectives. By utilizing the basic properties of the supply-demand matching process performed by grid operators, we develop a method to recover energy market's structure and predict the resulting nodal prices as a function of generation mix and system load on the grid. Our methodology uses the latest advancements in compressed sensing and statistics to cope with the high-dimensional and sparse power grid topologies, underlying physical laws, as well as scarce, public market data. Rigorous validations using Southwest Power Pool (SPP) market data demonstrate significantly higher accuracy of the proposed approach when compared to the state-of-the-art industry benchmark.

**Index Terms**—Locational Marginal Price (LMP), electricity price forecast, wholesale energy markets, statistical learning, big data, compressed sensing.

## I. INTRODUCTION

Development of distributed energy resource (DER) technologies enabled the owners of controllable energy assets to shape their wholesale market participation responsively and in a coordinated manner [1], [2]. To address the environmental and operational challenges, besides making the clean energy generation available and cheap, the question remains whether wholesale market prices could be inferred from the supply/demand mix on the grid and, then, used to create a feedback for "shaping" energy asset's production or consumption. In this paper we provide an affirmative response to this inquiry.

Power networks are defined by transmission lines that transport power from generators to loads. Generators and loads are connected to buses, which are commonly referred to as nodes of a power network. Following [3], many wholesale power markets in the US (CAISO, MISO, SPP, PJM, ISO New England, NYISO) adopted the concept of Locational Marginal Prices (LMPs) as

electricity prices at network (grid) nodes. The LMP at a specific node is defined as the marginal cost of supplying the next increment of load at that node, consistent with all power grid operating constraints. Even though the energy marketplace is organized differently on different continents, there is a global tendency to adopt the LMP approach. Various forms of such markets already exist in Singapore, New Zealand, Argentina (see a global market overview in [4]). On the other hand, European markets are zonal mainly due to the high connectivity of the European grid, where the intra-zonal congestions are managed using different, inefficient intra-zonal mechanisms, with a lot of ongoing discussion to move to nodal marginal pricing [5], [6]. In this paper, we introduce a methodology for predicting LMPs.

LMP markets are divided into day-ahead (DA) and real-time (RT, sometimes referred to as intra-day) markets. In a DA market, participants submit bids/offers to buy/sell energy. The Independent System Operator (ISO) then runs the Optimal Power Flow (OPF, [7]) program to derive DA LMPs for each grid's node. The OPF is an optimization problem that determines the generation schedule that minimizes the total system generation cost while satisfying demand/supply balance and network physical constraints [8]. Since DA scheduled supply may not meet real-time demand, ISOs also calculate RT LMPs every five minutes.

Our methodology is based on statistical learning techniques that take advantage of the sparsity properties induced by the nature of real grid topologies, underlying physical laws and the resulting OPF solution structure. The emerging field of statistical learning with sparsity [9] aims to utilize sparsity to help recover the underlying signal in a large set of data. Successful applications of sparse machine learning techniques include image/video processing [10], [11], pattern classification [12], face recognition [13], and customers preference learning [14]. In this paper, using the recent advancements in compressed sensing [15] and convex optimization [16], we utilize the OPF problem structure to infer the unknown grid topology, transmission line congestion regimes, and the resulting nodal prices as functions of grid-level

A. Radovanovic, and B. Chen are with Google, Inc. Sunnyvale, CA, 94089 (Email: [anaradovanovic@google.com](mailto:anaradovanovic@google.com), [bokanchen@google.com](mailto:bokanchen@google.com)) .

T. Nesti is with Centrum Wiskunde & Informatica (CWI), Amsterdam, Netherlands (Email: [T.Nesti@cwi.nl](mailto:T.Nesti@cwi.nl)). The underlying work in this paper was conducted during his internship at Google.

generation mix and load.

Applying the proposed methodology on SPP market data shows a significant improvement in prediction accuracy when compared to the state-of-art industry benchmark (Genscape [17]). Note that the proposed methodology assumes a decentralized, market participant-centric perspective, making it fully scalable. To the best of our knowledge, this is the first study that holistically incorporates the structural properties of the grid-level supply-demand matching (OPF), statistical inference and validation using the publicly available market data.

In the following two subsections we outline the basics of the power grid modeling, a version of the OPF formulation with its solution structure, as well as the key references in the domain.

#### A. Power grid model and the wholesale energy prices

The power grid is commonly modeled as a connected graph  $\mathcal{G} = \mathcal{G}(\mathcal{N}, \mathcal{L})$ , where the set of nodes  $\mathcal{N}$  represents  $n$  buses in the system and the set of edges  $\mathcal{L}$  model  $m$  transmission lines. We use  $\mathbf{g}, \mathbf{d} \in \mathbb{R}^n$  to denote the generation and demand vectors; use  $C_i(\cdot)$  to denote the cost function of generation at node  $i$ , which is typically modeled as an increasing quadratic function; and use  $\underline{\mathbf{g}}, \bar{\mathbf{g}} \in \mathbb{R}^n$  and  $\underline{\mathbf{f}}, \bar{\mathbf{f}} \in \mathbb{R}^m$  to denote the vectors of generation and transmission capacity limits. Then the OPF can be formulated as the following optimization problem:

$$\min_{\mathbf{g}} \sum_{i=1}^n C_i(g_i) \quad (1)$$

$$\text{s.t. } \mathbf{1}^T(\mathbf{g} - \mathbf{d}) = \mathbf{0}, \quad : \lambda \quad (2)$$

$$\underline{\mathbf{g}} \leq \mathbf{g} \leq \bar{\mathbf{g}}, \quad : \boldsymbol{\tau}^-, \boldsymbol{\tau}^+ \quad (3)$$

$$\underline{\mathbf{f}} \leq \mathbf{T}(\mathbf{g} - \mathbf{d}) \leq \bar{\mathbf{f}}, \quad : \boldsymbol{\mu}^-, \boldsymbol{\mu}^+, \quad (4)$$

where matrix  $\mathbf{T} \in \mathbb{R}^{m \times n}$ , known as the Power Transfer Distribution Factors matrix (PTDF), is used to map nodal generation and demand to active power flows over transmission lines under the widely used assumption of the direct current approximation [18]. The operators  $\leq$  and  $\geq$  are understood entry-wise. Following the notation and derivation in [19], the PTDF matrix can be written as

$$\mathbf{T} = [\mathbf{0} \ \mathbf{DAB}^{-1}], \quad (5)$$

where matrices  $\mathbf{D}, \mathbf{A}, \mathbf{B}$  describe topological and physical properties of the grid. In particular,  $\mathbf{A} \in \mathbb{R}^{m \times (n-1)}$  is the sub-matrix of the edge-node incidence matrix  $\tilde{\mathbf{A}}$  of  $\mathcal{G}$  obtained by deleting the first column, while  $\mathbf{B} \in \mathbb{R}^{(n-1) \times (n-1)}$  is the sub-matrix of the weighted Laplacian matrix of  $\mathcal{G}$  obtained by deleting the first row and the

first column. Finally,  $\mathbf{D} \in \mathbb{R}^{m \times m}$  is a diagonal matrix with  $\mathbf{D}_{\ell\ell} = x_\ell^{-1}$ , with  $x_\ell > 0$  denoting the reactance of line  $\ell \in \mathcal{L}$ . The reduced dimension from  $n$  to  $n-1$  stems from the nullity of the connected grid graph, i.e.  $\tilde{\mathbf{A}}\mathbf{1} = \mathbf{0}$ . In order to ensure the uniqueness of the optimal solution, without loss of generality, we remove from consideration the node corresponding to the first column, which is selected as the reference bus. In view of the definitions above, matrix  $\mathbf{A}$  is a full-column rank matrix, and  $\mathbf{B}$  is strictly positive definite with non-positive off-diagonal entries. The scalar  $\lambda$  and the vectors  $\boldsymbol{\tau}^+, \boldsymbol{\tau}^-, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-$  are the Lagrange multipliers of the corresponding equations. For more detailed derivation and discussion, an interested reader is referred to [19].

Potential generalizations of the optimization formulation above would involve additional operational constraints, such as ramping up/down constraints, power factor constraints, as well as treatments of the reactive power transfer and voltage variation bounds [20]. However, in this paper we show that, by utilizing the basic approximation in (1), we are able to capture the market structure and the dominant drivers of its dynamics. In addition, the recent advancements in power system technologies and changing regulations (e.g. [21]) will make the impact of reactive power transfer and the related voltage variations less exaggerated, and the DC-OPF approximation even more accurate.

LMPs are the shadow prices of the real power balance constraints of OPF [22]. More formally, they can be represented as

$$\text{LMP} = \frac{\partial \mathcal{L}}{\partial \mathbf{d}} = \lambda \mathbf{1} + \mathbf{T}^T \boldsymbol{\mu}, \quad (6)$$

where  $\frac{\partial \mathcal{L}}{\partial \mathbf{d}}$  denotes the partial derivative of the Lagrangian function of the OPF evaluated at the optimal solution, and  $\boldsymbol{\mu} = \boldsymbol{\mu}^- - \boldsymbol{\mu}^+$ . The entries of  $\boldsymbol{\mu}$  corresponding to uncongested lines ( $\underline{f}_\ell < f_\ell < \bar{f}_\ell$ ) are equal to zero, while the components corresponding to congested lines are different than zero (in particular,  $\mu_\ell^+ > 0$  iff  $f_\ell = \bar{f}_\ell$  and  $\mu_\ell^- < 0$  iff  $f_\ell = \underline{f}_\ell$ ). As a consequence, if there are no congested lines, all LMPs are equal, i.e.  $\text{LMP}_i = \lambda, \forall i \in \mathcal{N}$ , and the common value  $\lambda$  in (6) is called the marginal energy component (MEC). The energy component reflects the marginal cost of energy at the reference bus. On the other hand, if some lines are congested, we have  $\boldsymbol{\mu} \neq \mathbf{0}$  and, thus, the LMPs are different (see Figure 1); we call the second term  $\tilde{\boldsymbol{\pi}} = \mathbf{T}^T \boldsymbol{\mu}$  in (6) the marginal congestion component (MCC); in particular,  $\tilde{\boldsymbol{\pi}}_i$  reflects the marginal cost of congestion at bus  $i$  relative to the reference bus.

When ISOs calculate LMPs, they also include the loss component, which is related to the heat dissipated on

transmission lines, and is typically negligible compared to the other price components [23]. For this reason, we omit it from consideration in this paper, and end up with the marginal energy LMP component (same across all grid nodes) and marginal congestion LMP component as defined in the previous paragraph and expression (6).

If we recall the definition in (5), the marginal congestion price vector (excluding the reference bus) can be presented as

$$\boldsymbol{\pi} = \mathbf{B}^{-1} \mathbf{A}^T \mathbf{D} \boldsymbol{\mu} = \mathbf{B}^{-1} \mathbf{s} \in \mathbb{R}^{n-1}, \quad (7)$$

with  $\mathbf{s} = \mathbf{A}^T \mathbf{D} \boldsymbol{\mu}$ .  $\mathbf{s}$  contains the information on the congested lines since  $\mathbf{s} = \sum_{\ell=1}^m \mu_{\ell} x_{\ell}^{-1} \mathbf{a}_{\ell}$ , where  $\mathbf{a}_{\ell} \in \mathbb{R}^{n-1}$  is the  $\ell$ -th column of  $\mathbf{A}^T$ . Thus, by stacking historical  $\boldsymbol{\pi}, \mathbf{s}$  for  $T$  different time intervals as columns of the matrices  $\boldsymbol{\Pi}, \mathbf{S} \in \mathbb{R}^{(n-1) \times T}$ , we can rewrite (7) in matrix form as

$$\boldsymbol{\Pi} = \mathbf{B}^{-1} \mathbf{S}. \quad (8)$$

In Section IV, we use the previous relationship and the properties of matrices  $\mathbf{B}$  and  $\mathbf{S}$  to recover diverse congestion regimes that occur in a grid.

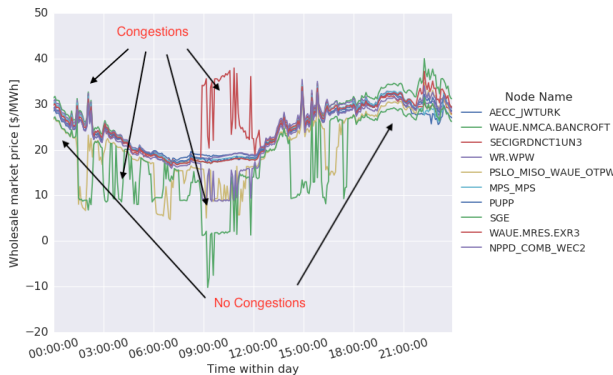


Fig. 1: Real-time market price for randomly selected SPP nodes.

### B. Related literature

The previously published work in the same domain can be split based on whether it takes an operator-centric or a participant-centric point of view. In the former case, the full knowledge of supply bids, grid topology, and physical properties of the network allow for the explicit computation of nodal LMPs as dual variables of the corresponding OPF optimization. The relevant papers mostly study the impact of uncertainty in total grid load or renewable generation on the resulting prices, while relying on the fact that changes in LMP regimes happen at the so called critical load

levels [24]–[28], whose number exponentially grows with the size of a grid, making the proposed approaches intractable for use in practice. The market participant-centric point of view has been much less addressed in the existing literature. [29] utilizes the structure of the OPF formulation to infer states of transmission lines using only the zonal load levels (no generation saturation is considered). Through the so called System Pattern Regions (SPR), zonal prices are obtained by learning the map between the zonal load and the corresponding zonal price, which introduces a large forecasting error. On the other hand, [30] presents a data-driven approach that exploits structural characteristics by learning nodal prices as a function of nodal loads using support vector machines (SVMs), for a synthetically generated, small grid example. In [31], the authors proposed an inverse optimization approach to estimate the parameters in the OPF, by assuming full knowledge of supply bids, nodal generation and prices, and then obtain price prediction by solving the OPF with new supply and demand data. Compared to the aforementioned approaches, our methods are more scalable. By adding consideration of generation mix, which strongly influences the market price, we are also able to produce nodal price forecasts that are more accurate than the industry benchmark, while only using publicly available data.

## II. ENERGY MARKET STRUCTURE

As Subsection I-A suggests, the nodal wholesale prices are functions of nodal demand and generation. Here, we state the key results from [29], [30], [32] which formalize the market structure using pricing regimes, represented via a vector of flags which indicate the marginal status of generators and congestion status of transmission lines at optimality. Our approach, discussed in Section IV, utilizes these theoretical concepts and parametrizes pricing regimes by a vector of publicly available grid-level generation mix and regional load (called system load) data.

For convenience, we reformulate the OPF problem defined by equations (1) - (4) as:

$$\min_{\mathbf{g}} \mathbf{J}_1^T \mathbf{g} + \frac{1}{2} \mathbf{g}^T \mathbf{J}_2 \mathbf{g} \quad \text{s.t.} \quad \mathbf{A} \mathbf{g} \leq \mathbf{b} + \mathbf{E} \boldsymbol{\theta}, \quad (9)$$

where  $\mathbf{g}$  are the optimization variables denoting the nodal generation,  $\boldsymbol{\theta} = [\mathbf{d} \ \bar{\mathbf{g}}]^T$  is a vector of nodal loads and generation capacities, and  $\mathbf{J}_1 \in \mathbb{R}^n$ ,  $\mathbf{J}_2 \in \mathbb{R}^{n \times n}$  define the linear and quadratic costs of generation, respectively. Similarly as in [32], here, we consider generation that has variable and fixed costs of production, but faces no startup, shutdown, no-load costs, or ramping constraints. Note that the linear and quadratic cost functions, as well as

the corresponding generation production range, constitute generators' bids. We assume that  $\mathbf{J}_2$  is diagonal positive definite. The matrices  $\mathbf{A}$ ,  $\mathbf{E}$  and the vector  $\mathbf{b}$  are given below:

$$\mathbf{A} = \begin{bmatrix} \mathbf{1}^T \\ -\mathbf{1}^T \\ \mathbf{T} \\ -\mathbf{T} \\ \mathbf{I} \\ -\mathbf{I} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \underline{\mathbf{f}} \\ \underline{\mathbf{f}} \\ \mathbf{0} \\ -\underline{\mathbf{g}} \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} \mathbf{1}^T & \mathbf{0}^T \\ -\mathbf{1}^T & \mathbf{0}^T \\ \mathbf{T} & \mathbf{0} \\ -\mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$\boldsymbol{\theta} = \begin{bmatrix} \underline{\mathbf{d}} \\ \underline{\mathbf{g}} \end{bmatrix} \quad (10)$$

As discussed in [32], the feasible region of the optimization problem in (9) is convex, compact and polyhedral, thus a polytope. Its facets correspond to the so called pricing regimes uniquely defined by the set of marginal generators and congested transmission lines. To that end, if  $\mathcal{J}$  denote the index set of constraints of (9), we define

$$\mathcal{B}(\boldsymbol{\theta}) = \{i \in \mathcal{J} \mid A_i \mathbf{g}^* = \mathbf{b} + E_i \boldsymbol{\theta}\}$$

$$\mathcal{B}^{\mathcal{C}}(\boldsymbol{\theta}) = \{i \in \mathcal{J} \mid A_i \mathbf{g}^* < \mathbf{b} + E_i \boldsymbol{\theta}\}$$

The set  $\mathcal{B}$  corresponds to binding (active) constraints, while  $\mathcal{B}^{\mathcal{C}}$  correspond to non-binding constraints. Clearly,  $\mathcal{B} \cap \mathcal{B}^{\mathcal{C}} = \emptyset$  and  $\mathcal{B} \cup \mathcal{B}^{\mathcal{C}} = \mathcal{J}$ . We identify the pricing regime with the corresponding set of binding constraints  $\mathcal{B}$ . Following [29] (Proposition 1), and the derivation in Section III of [32], we have the following results, which constitute the theoretical foundations of the methodology presented in this paper:

**Theorem II.1.** *Assume that the OPF problem (9) is non-degenerate<sup>1</sup>. Then:*

- 1) *Its feasible region can be uniquely partitioned into disjoint open convex polytopes uniquely defined by  $\mathcal{B}$ ;*
- 2) *Within each pricing regime  $\mathcal{B}$ , the optimal generation  $\mathbf{g}^*$  and the associated vector of LMPs, (6), are uniquely defined affine functions of  $\underline{\mathbf{d}}$  and  $\underline{\mathbf{g}}^*$ . Overall, the vector of LMPs over the whole feasible region is a continuous, piecewise affine function of nodal demand  $\underline{\mathbf{d}}$  and optimal generation vectors  $\underline{\mathbf{g}}^*$ .*

### III. INPUT MARKET DATA

The publicly available market data depends on the specific market and commonly includes historical grid-level generation mix, regional (system) load, and nodal

<sup>1</sup>See section 4.1.1 in [33] for the definition

LMPs. In case of the SPP market, historical generation mix is recorded at 5 minute time granularity and equals to the total average power produced across different types of generation (coal, natural gas, wind, solar, nuclear, etc.). System load mix comprises of regionally aggregated average demand values recorded at hourly time granularity, with 16 different load zones (see Figure 2).

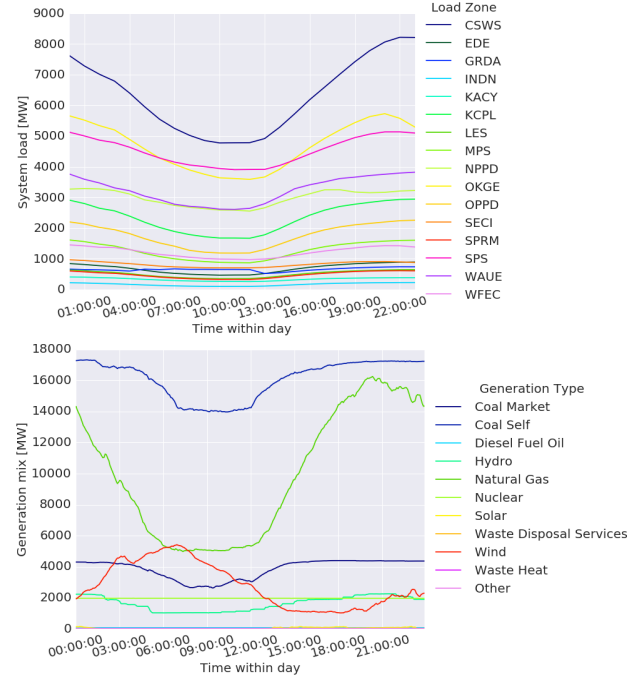


Fig. 2: Grid level supply and demand: system load (top), generation mix (bottom).

In addition to the grid level mix and the regional load data, operators release the corresponding real-time nodal prices at 5 min time granularity. For the purpose of training and validation, in this paper we use six months of SPP data, from June to November of 2017, with 929 nodes, where we exclude a few new nodal connections with limited price history. To facilitate the analysis, we scale the available generation and load data, and refer to the scaled data as  $\mathcal{M}$ -vectors defined as follows.

**Definition III.1.** *At any real-time interval  $t$ ,  $\mathcal{M}(t)$  refers to the vector of concatenated normalized generation mix, normalized system load, and scaled total demand, i.e., where the normalized generation of type  $\gamma$  equals to  $\sum_{i \in S_\gamma} g_i(t) / \sum_\gamma \sum_{i \in S_\gamma} g_i(t)$  ( $S_\gamma$  being equal to the set of generators of type  $\gamma$ ), system load  $\lambda$ 's normalized demand equals to  $\sum_{i \in S_\lambda} d_i(t) / \sum_\lambda \sum_{i \in S_\lambda} d_i(t)$  ( $S_\lambda$  being equal to the set of load nodes belonging to the system load  $\lambda$ ), and*

the scaled total load entry of the  $\mathcal{M}(t)$  vector equals to  $\frac{\sum_{\lambda} \sum_{i \in S_{\lambda}} d_i(t)}{\text{Average}_{t \in \mathcal{I}} [\sum_{\lambda} \sum_{i \in S_{\lambda}} d_i(t)]}$  (Average $_{t \in \mathcal{I}}$  is used to denote the sample average of the total, grid level, demand across all available time instances  $\mathcal{I}$ ).

Even though  $\mathcal{M}$ -vectors are time-indexed, for the reasons of simplicity, in the rest of the paper we omit time index when referring to  $\mathcal{M}(t)$ .

#### IV. NEW PRICE PREDICTION METHODOLOGY

In this section we propose our price prediction methodology that assumes no information on generators' placement, capacities and pricing curves, as well as grid topology, line capacities and load distribution across its nodes. We reconstruct the pricing regimes (defined in Section II) by first recovering its structure (Subsections IV-A, IV-B, and IV-C) and, then, utilizing the assumption that system loads and generators of the same type change synchronously within the same  $\mathcal{M}$ -regime (see Subsection IV-D). A summary of the methodology is in Figure 3.

The following conjectures are validated using the historical SPP data.

**Conjecture IV.1.** (i)  $\mathcal{M}$ -vectors can be classified into a small number of regimes (4 in the case of SPP market), to be referred to as  $\mathcal{M}$ -regimes (Subsection IV-A). (ii) Blind matrix recovery consistently (in time) reveals the same grid topology and congestion regimes, where each  $\mathcal{M}$ -regime commonly corresponds to a single dominant congestion regime (Subsections IV-B and IV-C).

The discussion in Section II provides theoretical concepts for piecewise linear relationships between LMPs and pricing regimes that are uniquely defined by nodal demands and dispatched generation. In order to relate nodal generation and load to the corresponding grid and region level quantities within each  $\mathcal{M}$ -regime, we make the following assumption.

**Assumption IV.1.** Within each  $\mathcal{M}$ -regime, all generators of the same type (e.g., wind, natural gas, etc.) preserve their production fraction with respect to the total grid level generation of the same type. Similarly, all load within the same geographic region preserves the same consumption ratio when compared to the total load in the region.

Intuitively, Assumption IV.1 treats a grid-level increase in, for example, wind supply, proportionally equal across all wind plants on the specific grid. Note that the constant ratio, in general, changes from one  $\mathcal{M}$ -regime to another within a day. This assumption enables us to extend the

piecewise affinity in Theorem II.1, and parametrize the pricing regimes using  $\mathcal{M}$ -vectors. Note that it cannot be validated given that the local, nodal, load and generation data is not available. However, it enables us to utilize the piecewise affinity and continuity across disjoint convex polytopes to efficiently fit the Multivariate Adaptive Regression Splines (MARS) models [34], [35] and recover nodal LMP vector as a function of  $\mathcal{M}$ -vectors.

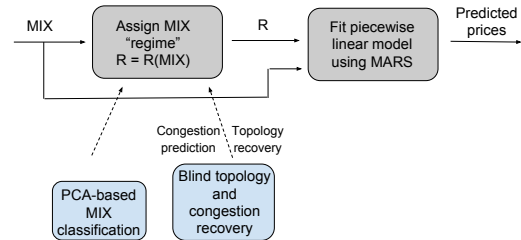


Fig. 3: Price prediction pipeline.

Next, we separately describe each of the modeling components and our approach in validating them.

##### A. Clustering into $\mathcal{M}$ regimes

We classify  $\mathcal{M}$ -vectors by first applying Principal Component Analysis (PCA, [36]) using the time indexed  $\mathcal{M}$ -vectors as defined earlier. Then, we perform k-means clustering [37] using the obtained lower dimensional  $\mathcal{M}$  representations, where, by applying the fairly standard elbow method, we end up with 4  $\mathcal{M}$ -regimes. The PCA revealed that only 4 dominant principal components explain 98% of its variance (see Figure 4). Interestingly, the same property is preserved across different time horizons.

##### B. Topology recovery

Using the structural properties discussed in Subsection I-A, we perform the blind matrix factorization to recover the congestion matrix  $\mathbf{S}$ .  $\mathbf{B}$  and  $\mathbf{S}$  enjoy the following structural properties: (i)  $\mathbf{B}$  is a positive definite M-matrix and is sparse, and (ii)  $\mathbf{S}$  is sparse and low-rank. The sparsity of  $\mathbf{B}$  follows from the fact that the graph underlying a power grid is usually weakly connected (mainly holds for grids in the USA, [38]). The fact the  $\mathbf{S}$  is sparse and low-rank follows from its definition (7) and the fact that almost always only a very small subset of transmission lines gets congested.



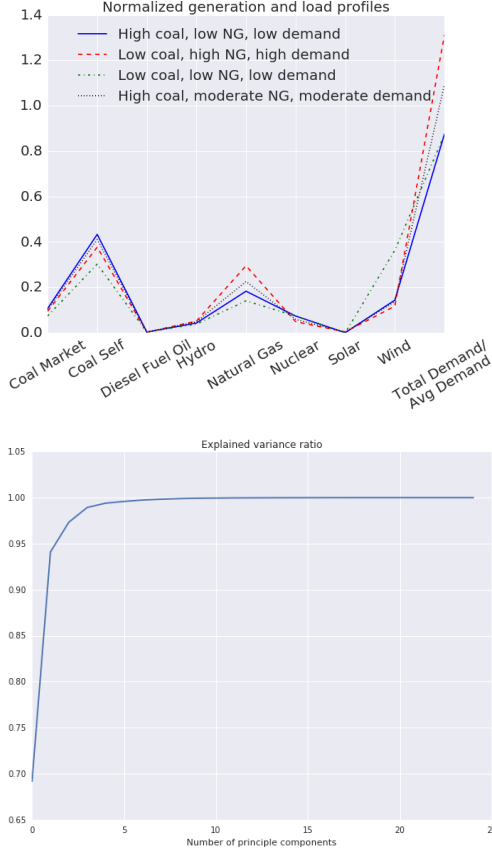


Fig. 4: Generation mix corresponding to the centroids of the PCA-based clusters (top), and explained ratio by the first four principal components (bottom). NG in the legend (top) is used to refer to Natural Gas.

We use the approach in [19] where matrices  $\mathbf{B}$  and  $\mathbf{S}$  are obtained by solving the following convex relaxation

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{S}} \quad & \|\mathbf{S}\|_1 + \kappa_1 \text{tr}(\mathbf{P}\mathbf{B}) - \kappa_2 \log |\mathbf{B}| \\ \text{s.t.} \quad & \mathbf{B}\mathbf{\Pi} = \mathbf{S}, \mathbf{B} \in \mathcal{C}, \end{aligned} \quad (11)$$

with  $\mathbf{P} = \mathbf{I} - \mathbf{1}\mathbf{1}^T$ ,  $\mathcal{C} := \{\mathbf{B} : \mathbf{B} \succeq \mathbf{0}, \mathbf{B} \leq \mathbf{I}\}$ ,  $\kappa_1, \kappa_2 \geq 0$ , and  $\mathbf{B} \succeq \mathbf{0}$  denoting a positive semidefinite matrix. The  $\leq$  operator is understood entry-wise. Given that the previous semidefinite program is hard to solve for large grids (order of  $\sim 1000x$  nodes), we utilize the Alternating Direction Methods of Multipliers (ADMM, [39]) to solve the program iteratively. Following [19], we first replace  $\mathbf{B}$  with three copies  $\mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \mathbf{B}^{(3)}$ , yielding an equivalent formulation of (11), and then define the matrices  $\mathbf{M}_{12}, \mathbf{M}_{13}, \mathbf{M}$  to be the Lagrange multipliers corresponding to the equality constraints of this new formulation.

Every iteration of the ADMM consists of three steps, during which the variables and Lagrange multipliers are updated by solving appropriate optimization problems, which are expressed in terms of the solutions computed at the previous steps and iterations. Leveraging the existence of closed form solutions for these optimization problems, the  $i$ th iteration of the ADMM reads:

$$\begin{aligned} \mathbf{B}^{(1),i+1} &= (\mathbf{B}^{(2),i} - \mathbf{M}_{12}^i + \mathbf{B}^{(3),i} - \mathbf{M}_{13}^i + \\ & \quad (\mathbf{S} - \mathbf{M})\mathbf{\Pi}^T - \frac{\kappa_1}{\rho}\mathbf{P})(2\mathbf{I} + \mathbf{\Pi}\mathbf{\Pi}^T)^{-1}, \\ \mathbf{B}^{(2),i+1} &= \min(\mathbf{B}^{(1),i+1} + \mathbf{M}_{12}^i, \mathbf{I}), \\ \mathbf{B}^{(3),i+1} &= \mathbf{U} \text{diag} \left( \left\{ \frac{1}{2} \left( s_k + \sqrt{s_k^2 + 4 \frac{\kappa_2}{\rho}} \right) \right\} \right) \mathbf{U}^T, \\ \mathbf{S}^{i+1} &= (\mathbf{B}^{(1),i+1} \mathbf{\Pi} + \mathbf{M}^i) \odot \mathbf{Y}^{(i+1)}, \\ \mathbf{M}_{12}^{i+1} &= \mathbf{M}_{12}^i + \rho(\mathbf{B}^{(1),i+1} - \mathbf{B}^{(2),i+1}), \\ \mathbf{M}_{13}^{i+1} &= \mathbf{M}_{13}^i + \rho(\mathbf{B}^{(1),i+1} - \mathbf{B}^{(3),i+1}), \\ \mathbf{M}^{i+1} &= \mathbf{M}^i + \rho(\mathbf{B}^{(1),i+1} \mathbf{\Pi} - \mathbf{S}^{i+1}); \end{aligned} \quad (12)$$

where  $\mathbf{B}^{(1),0} = \mathbf{B}^{(2),0} = \mathbf{B}^{(3),0} = \mathbf{I}$ ,  $\mathbf{S}^0 = \mathbf{B}^{(1),0} \mathbf{\Pi}$ ,  $\mathbf{M}_{12}^0 = \mathbf{M}_{13}^0 = \mathbf{0}$ ,  $\mathbf{M}^0 = \mathbf{0}$ ,  $s, \mathbf{U}$  are the eigenvalues and eigenvectors obtained through eigen decomposition of  $0.5 \cdot (\mathbf{B}^{(1),i+1} + \mathbf{M}_{12}^i)(\mathbf{B}^{(1),i+1} + \mathbf{M}_{12}^i)^T$ , and  $\mathbf{Y}^{(i+1)}$  is defined entry-wise as  $(\mathbf{Y}^{(i+1)})_{kh} = \max\left(0, 1 - \frac{\kappa_2}{\rho(|\mathbf{B}^{(1),i+1} \mathbf{\Pi} + \mathbf{M}^i|_{kh})}\right)$ . The symbol  $\odot$  denotes the Hadamard, or entry-wise, product, and the minimum operator is understood entry-wise.

In practice, nodal connections slowly change due to sporadic repairs and new nodes and, therefore, we expect  $\mathbf{B}$  to be approximately constant. We validate this hypothesis by taking 7 consecutive weeks of real time market prices and, for each of them, we run the matrix recovery algorithm to infer  $\mathbf{B}_{w,1}, \mathbf{B}_{w,2}, \mathbf{B}_{w,3}, \dots, \mathbf{B}_{w,7}$ . To evaluate the difference in the recovered links, we first perform entry-wise normalization of all the recovered matrices by dividing each entry with the entry-wise maximum absolute value to obtain scaled matrices  $\hat{\mathbf{B}}_{w,1}, \hat{\mathbf{B}}_{w,2}, \hat{\mathbf{B}}_{w,3}, \dots, \hat{\mathbf{B}}_{w,7}$ . Then, we count the identified links by counting off-diagonal entries with absolute values exceeding some given threshold value.

The result of counting the identified grid links for a given threshold across all recovered matrices exhibit a surprising proximity (see Figure 5), despite the variable impact of the numerical precision criteria of the blind recovery algorithm, as well as changing link reactances due to weather conditions and variations in heating induced by the energy transfer. In Table I, we use  $p(\mathbf{B}_{w,i-1}, \mathbf{B}_{w,i})$  to express the percent of links identified from week  $i$ th matrix  $\mathbf{B}_{w,i}$ , that are not recovered by week  $(i-1)$ th matrix  $\mathbf{B}_{w,i-1}$ .

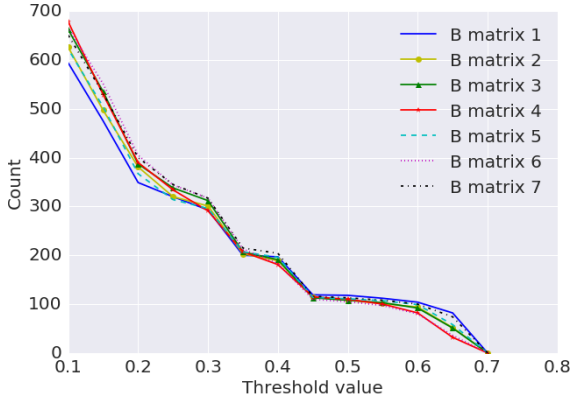


Fig. 5: Number of identified transmission lines corresponding to the dominant  $\hat{\mathbf{B}}$ -entry values.

TABLE I: Differences in recovered links.

	Percent
$p(\mathbf{B}_{w,1}, \mathbf{B}_{w,2})$	6%
$p(\mathbf{B}_{w,2}, \mathbf{B}_{w,3})$	6%
$p(\mathbf{B}_{w,3}, \mathbf{B}_{w,4})$	6%
$p(\mathbf{B}_{w,4}, \mathbf{B}_{w,5})$	4%
$p(\mathbf{B}_{w,5}, \mathbf{B}_{w,6})$	6%
$p(\mathbf{B}_{w,6}, \mathbf{B}_{w,7})$	3%

### C. Congestion regime recovery

Apart from the topology matrix  $\mathbf{B}$ , the blind matrix recovery algorithm recovers the congestion matrix  $\mathbf{S}$  (see Figure 6). In contrast to [19], in the present work we are mostly interested in recovery of the congestion regimes (i.e. collections of congested lines) within each  $\mathcal{M}$ -regime obtained by the previously discussed PCA and clustering. Surprisingly, by clustering columns of the recovered matrices  $\mathbf{S}$ , each  $\mathcal{M}$ -regime ends up having one dominant congestion regime. By using multinomial logistic regression classification, we map the deviation from a typical  $\mathcal{M}$ -vector within each  $\mathcal{M}$ -regime to a congestion cluster. Misclassification happens in less than 4% of instances across all  $\mathcal{M}$ -regimes and, commonly, corresponds to extremely large price spikes (bursts) in real time price, which possibly is a result of the limited information used in this approach (only grid level generation and system load).

### D. Performance analysis

We analyze the performance of the proposed methodology in case of two SPP nodes: SPPNORTH and SPPSOUTH hub (note that the proposed approach allows us to predict the price at any node on the grid). The

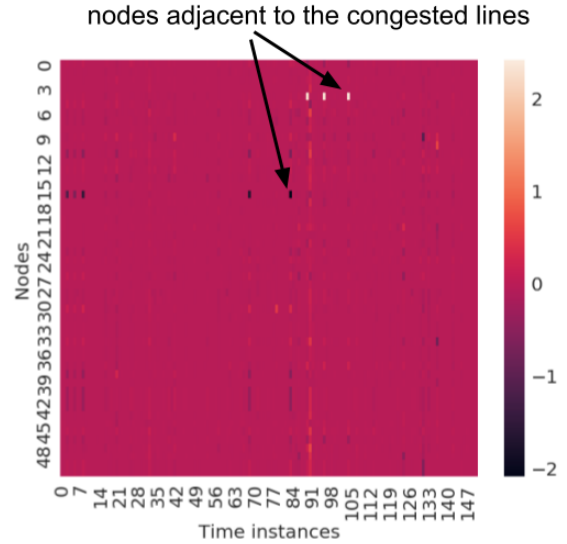


Fig. 6: Heatmap of a segment of the congestion matrix.

training dataset spans only 7 weeks, from June 12th to July 30th 2017, while the testing period lasts for the following 2 weeks, from July 31st to August 13th 2017. Our validation process consists of:

- 1) Initial clustering into  $\mathcal{M}$ -regimes using the training period's data, which then is applied to classify the  $\mathcal{M}$ -vectors from the next 2 weeks of the testing period;
- 2) Blind matrix recovery using only a week of the historical price data, from July 24th 2017 to July 30th 2017. The recovered congestion matrix is used to identify congestion regimes.
- 3) Day-ahead prediction of the RT market prices for a given node using the MARS models trained for each congestion regime<sup>2</sup> separately, where the training data consists of the  $\mathcal{M}$ -vectors and the corresponding RT prices from all days prior to the test day.

The predictive performance in the testing period is evaluated using the Median of the Absolute Percent Error (median is used to eliminate the impact of the extremely large price spikes in the real time prices that commonly happen due to unpredictable infrastructure failures/changes). The state-of-the-art baseline predictions for the observed nodes are purchased from Genscape [17], which is commonly referred to as the industry benchmark.

We evaluate the predictive performance of the new,

<sup>2</sup>As discussed in Subsection IV-C, we infer a single, dominant congestion (sub)regime within each  $\mathcal{M}$ -regime and, thus, we end up training MARS models for each  $\mathcal{M}$ -regime as a whole.

proposed, approach for the following two scenarios: (i) with perfect knowledge of  $\mathcal{M}$ , (ii) using day-ahead  $\mathcal{M}$  forecasts obtained from the day-ahead forecasts for generation mix and system load purchased from Tomorrow [40]. The performance comparisons are summarized in Table II. Our validations show that the new methodology significantly outperforms the industry benchmark, where the performance gap can be controlled by the forecasting accuracy of the grid level  $\mathcal{M}$ .

TABLE II: Median absolute percent error on the test set.

Approach	SPPSOUTH Hub	SPPNORTH Hub
Benchmark	30%	32%
New (actual $\mathcal{M}$ )	16%	20%
New ( $\mathcal{M}$ forecast)	25%	24%

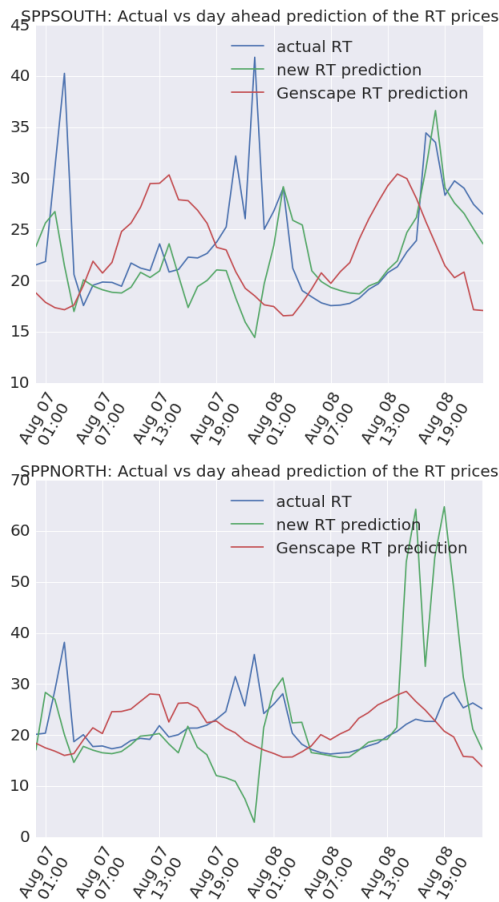


Fig. 7: Intraday comparison between the actual and forecasted RT hourly prices, actual  $\mathcal{M}$  scenario.

A typical intraday samples of the day ahead RT price predictions and the corresponding actuals are captured

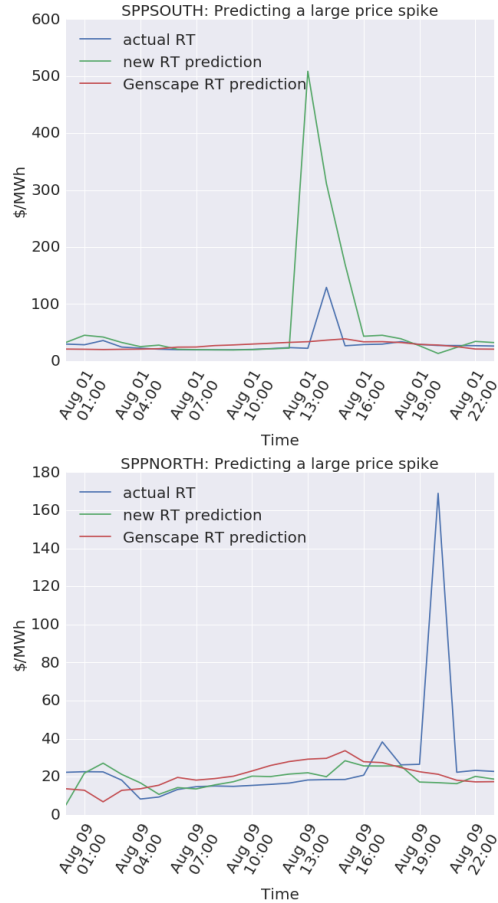


Fig. 8: A case of the predicted spike (top) and unpredicted spike (bottom) in RT price, actual  $\mathcal{M}$  scenario.

in Figure 7. Even though the new approach tends to capture the somewhat spiky behavior of the RT prices, in some cases of large spikes, it fails to do so (see the bottom figure in Figure 8). The most probable reason might be a sudden infrastructure failure or repair, the conclusion being that not all the spikes are predictable from changes in the generation and system load mix. Furthermore, our approach utilizes only a simple version of the DC-OPF formulation, without taking into account reserves and ramping constraints, which can be a possible cause of the market price spikes. Since the process of supply-demand matching for DA energy market involves solving the same optimization, the models trained for the RT price prediction can be used to infer DA prices, while the inputs in this case would correspond to the day-ahead cleared generation and system load, and is typically less variable (i.e., we can expect smaller prediction errors).

Finally, the distribution of the absolute percent errors at different instances in time is presented in Figure 9.



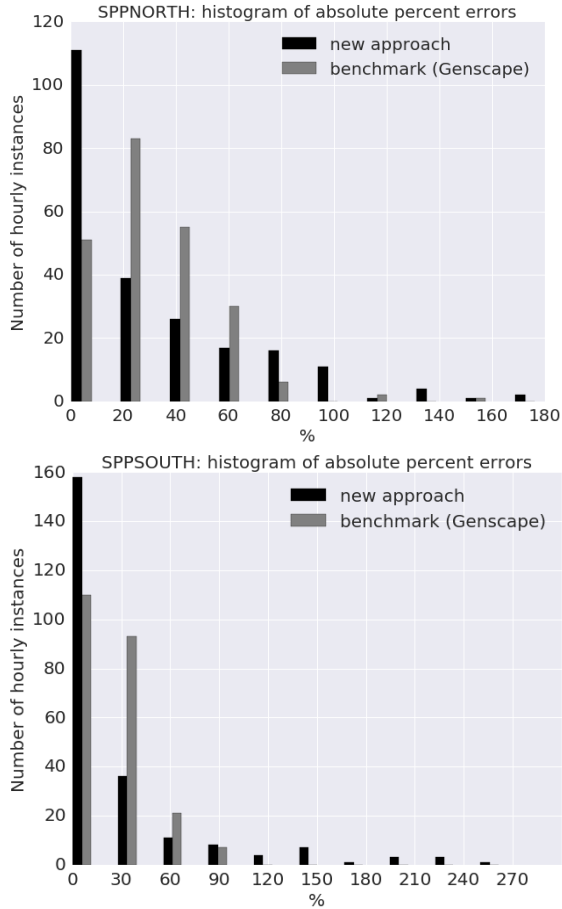


Fig. 9: Histogram of the Absolute Percent Errors at the test instances.

## V. CONCLUSIONS AND FUTURE WORK

In this paper we show that the wholesale energy market structure can be inferred using very limited, publicly available, historical market data (grid level generation type mix, system load mix, and nodal prices). By utilizing the basic underlying physical model that captures generation-load matching on the grid, we present a day ahead nodal price prediction methodology that, under a rigorous validation framework, significantly outperforms the industry benchmark (Genscape [17]).

The proposed methodology can be enhanced with additional data. For example, the statistical models in this paper did not capture seasonality patterns in generation and load, as well as the impact of other relevant data, such as available reserves and their prices, or ramping constraints. To that end, we believe that using the additional data sources can only improve the prediction performance, which will be part of our future

investigations. The other potential area of exploration involves studying the impact of localized measurements at market participants' sites to improve the corresponding local market predictions. After all, it is expected that each market participant's goal is to maximize its own financial reward and environmental impact and, therefore, improving market predictions is the key instrument for achieving these objectives.

## ACKNOWLEDGMENT

The authors would like to thank Thomas Olavson and John Platt for their support and constructive comments.

## REFERENCES

- [1] C. Eid, P. Codani, Y. Perez, J. Reneses, and R. Hakvoort, "Managing electric flexibility from distributed energy resources: A review of incentives for market design," *Renewable and Sustainable Energy Reviews*, vol. 64, pp. 237 – 247, 2016.
- [2] I. Pavić, M. Beus, H. Pandžić, T. Capuder, and I. Štritof, "Electricity markets overview; market participation possibilities for renewable and distributed energy resources," in *2017 14th International Conference on the European Energy Market (EEM)*, June 2017, pp. 1–5.
- [3] FERC, "White paper wholesale power market platform," April 2003.
- [4] F. E. International Transmission Pricing Review, <https://www.ea.govt.nz/dmsdocument/2539>, [Online].
- [5] K. Neuhoff, B. F. Hobbs, and D. M. Newbery, "Congestion management in european power networks: criteria to assess the available options," 2011.
- [6] N. P. P. in Poland, [https://sites.hks.harvard.edu/hepg/Papers/2011/IAEE\\_TSIKORSKI\(v1%200\).pdf](https://sites.hks.harvard.edu/hepg/Papers/2011/IAEE_TSIKORSKI(v1%200).pdf), [Online].
- [7] M. Huneault and F. D. Galiana, "A survey of the optimal power flow literature," *IEEE Transactions on Power Systems*, vol. 6, no. 2, pp. 762–770, May 1991.
- [8] F. Schweppe, M. Caramanis, and R. Tabors, "Evaluation of Spot Price Based Electricity Rates," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-104, no. 7, pp. 1644–1655, jul 1985.
- [9] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.
- [10] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [11] J. Mairal, G. Sapiro, and M. Elad, "Learning multiscale sparse representations for image and video restoration," *Multiscale Modeling & Simulation*, vol. 7, no. 1, pp. 214–241, 2008.
- [12] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 609–616.
- [13] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.
- [14] V. F. Farias and A. A. Li, "Learning preferences with side information," *Management Science*, p. to appear, 2017.
- [15] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.
- [16] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE transactions on information theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

- [17] Genscape, <https://www.genscape.com/>.
- [18] K. Purchala, L. Meeus, D. V. Dommelen, and R. Belmans, "Usefulness of dc power flow for active power flow analysis," in *IEEE Power Engineering Society General Meeting, 2005*, June 2005, pp. 454–459 Vol. 1.
- [19] V. Kekatos, G. B. Giannakis, and R. Baldick, "Online energy price matrix factorization for power grid topology tracking," *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1239–1248, May 2016.
- [20] D. Mehta, D. K. Molzahn, and K. Turitsyn, "Recent advances in computational methods for the power flow equations," in *2016 American Control Conference (ACC)*, July 2016, pp. 1753–1765.
- [21] C. P. U. Commission, <http://www.cpuc.ca.gov/Rule21/>, [Online].
- [22] T. Orfanogianni and G. Gross, "A general formulation for lmp evaluation," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1163–1173, Aug 2007.
- [23] S. P. Pool, "2016 annual state of the market report," [https://www.spp.org/documents/53549/spp\\_mmu\\_asom\\_2016.pdf](https://www.spp.org/documents/53549/spp_mmu_asom_2016.pdf), [Online].
- [24] R. Bo and F. Li, "Probabilistic lmp forecasting considering load uncertainty," *IEEE Transactions on Power Systems*, vol. 24, no. 3, pp. 1279–1289, Aug 2009.
- [25] F. Li and R. Bo, "Congestion and price prediction under load variation," *IEEE Transactions on Power Systems*, vol. 24, no. 2, pp. 911–922, May 2009.
- [26] R. Bo and F. Li, "Probabilistic lmp forecasting under ac optimal power flow framework: Theory and applications," *Electric Power Systems Research*, vol. 88, no. Supplement C, pp. 16 – 24, 2012.
- [27] Y. Ji, R. J. Thomas, and L. Tong, "Probabilistic forecasting of real-time lmp and network congestion," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 831–841, March 2017.
- [28] W. Deng, Y. Ji, and L. Tong, "Probabilistic Forecasting and Simulation of Electricity Markets via Online Dictionary Learning," *ArXiv e-prints*, Jun. 2016.
- [29] Q. Zhou, L. Tesfatsion, and C. C. Liu, "Short-term congestion forecasting in wholesale power markets," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2185–2196, Nov 2011.
- [30] X. Geng and L. Xie, "Learning the lmp-load coupling from data: A support vector machine based approach," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1127–1138, March 2017.
- [31] J. R. Birge, A. Hortaçsu, and J. M. Pavlin, "Inverse optimization for the recovery of market structure from market outcomes: An application to the miso electricity market," *Operations Research*, vol. 65, no. 4, pp. 837–855, 2017.
- [32] J. Mather, E. Baeyens, D. Kalathil, and K. Poolla, "The geometry of locational marginal prices," submitted to *IEEE Transactions on Power Systems*.
- [33] A. Bemporad, M. Morari, V. Dua, and E. N. Pistikopoulos, "The explicit linear quadratic regulator for constrained systems," *Automatica*, vol. 38, no. 1, pp. 3 – 20, 2002.
- [34] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.
- [35] M. A. R. Splines, <https://cran.r-project.org/web/packages/earth/index.html>, [Online].
- [36] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 498–520, 1933.
- [37] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.
- [38] M. Babakmehr, M. G. Simões, M. B. Wakin, and F. Harirchi, "Compressive sensing-based topology identification for smart grids," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 532–543, April 2016.
- [39] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.
- [40] Tomorrow, <http://www.tmrow.com/>, <http://www.electricitymap.org>, [Online].