Database Development and Analysis of PJM Energy Markets

David Froot
ESE 499
Systems Engineering
Concentration in Economics
Senior Design Project
4/27/12

Advisor:
Dennis Flaherty, CEO
Great Bay Energy LLC.
dflaherty@greatbayenergy.com
AM Tower, 9th Floor
207 Calle Del Parque
San Juan, PR 00912

Abstract
Congestion pricing is an important source of pricing variability in wholesale electricity markets, and its behavior is often highly volatile. Though Day-Ahead prices for energy are calculated with the best available data, they often misestimate the costs in real time, and result in risk for market participants. In conjunction with Great Bay Energy LLC, this project will build a database to facilitate future market research using MATLAB software, and build a variety of analysis tools in order to examine the behavior of congestion spread pricing.
Background

PJM is a regional transmission organization (RTO) that coordinates the movement and transfer of electricity in thirteen states in the Northeast and Midwest. It acts as a neutral party that operates wholesale electric markets with energy producers and energy consumers, interacting through PJM to buy and sell energy. PJM runs a two-market system, with a Day-Ahead (DA) market and Real-Time (RT) market for electricity.

The Day-Ahead market is formed in part because energy producers need to schedule the operation of their machinery. Producers need to know in advance when and where their energy will be needed. PJM calculates, based on past data, the best estimate for the demand at each hour for the next day. It matches bids from producers and consumers to ensure energy demands are met. Because energy is produced and consumed in the same instant, load balancing is crucial; even a small deviation in the power on the grid can cause a blackout.

However, because in the real world people behavior unpredictably, load forecasting is never perfect. This imbalance between scheduled and real demand requires the formation of an additional, Real-Time market. If demand is higher than predicted, PJM needs more energy on the grid, and thus prices rise. Producers see these rising prices, and respond by deploying available resources. Likewise, consumers respond by decreasing energy consumption. While most standard residential consumers do not face real time energy prices, some consumers such as factories and other power intensive customers do pay these prices. Prices act as economic signals to balance the grid load.

Prices are set for each of the 10,000 pricing locations, in a system known as Locational Marginal Pricing (LMP). These prices are based on the cost of energy production, loss prices due to transmission, and the cost of moving the electricity across the grid (if necessary), known
as congestion pricing. This price arises from the fact that only a limited amount of power can be transported on an electrical wire, so transmission capacity becomes a monetized commodity. PJM offers market participants a variety of financial instruments to manage the price risks of the market, however this paper will focus on a specific instrument known as the Up-to Congestion product. This spread product allows market participants to financially move power from one point to another within the PJM footprint.

Day-Ahead spreads are defined as the difference between the congestion and loss prices at any node A, and the congestion and loss prices at any other node B. Market participants bid on this Day-Ahead spread price. Real-Time spreads are formulated in the same way, and the contracts are settled in the Real-Time market. Day-Ahead/Real-Time (DART) spreads are defined as the difference between Day-Ahead and Real-Time spreads. The Up-to Congestion spread product is only available for approximately 500 nodes on the PJM grid.

The Day-Ahead market operates as a blind Dutch auction. Participants submit their bids into the Day-Ahead market by 12:00 PM Eastern Standard Time, and PJM determines which bids clear the market by 4:00 PM by using the least-cost configuration for producers and consumers. Given that the market functions as a blind auction, there is no price discovery as in typical markets. Thus, deciding how to bid for energy is more difficult than in markets where the market price is visible to all. Understanding how to better bid on spreads is an important issue for market participants, and of interest to Great Bay Energy. Additional software is needed to research and analyze the spread data, including software to compile the large quantity of available data into a format that lends itself readily to research in a MATLAB environment. MATLAB will allow Great Bay Energy employees to utilize built-in functions to better analyze market-pricing data, and constantly have the most up-to-date data available.
Problem Statement

This project, in conjunction with Great Bay Energy, will analyze pricing data on the PJM electric grid, and develop related software for the future research of spread data. Specifically, the question of whether or not DART spreads are a leading or lagging indicator of Day-Ahead spreads will be examined.

To answer the problem statement, there are two distinct phases of the project. Before analysis can begin, software is necessary to analyze the pricing data for the previous three years. This requires a variety of functions, built in MATLAB, which can load and select the relevant data from available sources into a single database. Given the size of the database, this must be constructed efficiently, and contain functionality for adding more data to the database as it is generated from the marketplace. Other software is needed to better visualize, understand and work with the data involved. Once the data management applications are built, the next phase involves the analysis laid out in the original problem statement. Using the data software, as well as wide variety of analysis tools built in MATLAB, this project will attempt to better understand the behavior and relationship between DART and Day-Ahead spreads.

Term Definitions:

DA Price = DA Congestion Price + DA Loss Price

RT Price = RT Congestion Price + RT Loss Price

DA Spread = DA Price at Node A – DA Price at Node B

RT Spread = RT Price at Node A – DA price at Node B

DART Spread = RT Spread – DA spread

Intra-Node DART spread = RT Price at Node A – DA Price at Node A
Database Generation

Given that each node has six pricing points for every hour, this database must hold 1.5 billion pricing points in order to store three years worth of data, and constantly grows as new data is added. This project created several programs to retrieve information from online sources of pricing data. This is then compiled into a large database containing all pricing points from all nodes.

In both the DA and RT market, there is a congestion price, loss price, and total Locational Marginal Price. Up-to Congestion bids focus only on the congestion and loss prices. The number of nodes on the system changes, so importing data must account for missing or new data, in order to maximize the amount of data extracted. Using import software built in MATLAB, we generated two databases, which most efficiently store data given the intended uses.

Data Management Graphical User Interface

Because new data is generated from the PJM interconnection daily, a MATLAB graphical user interface (GUI) is needed to make adding new data into the existing database easy. This software loads each new set of data, and stores the spreads which are defined as follows:

\[
S^{\text{DA}} = (C^{\text{DA}}_A + L^{\text{DA}}_A) - (C^{\text{DA}}_B + L^{\text{DA}}_B)
\]
\[
S^{\text{RT}} = (C^{\text{RT}}_A + L^{\text{RT}}_A) - (C^{\text{RT}}_B + L^{\text{RT}}_B)
\]
\[
S^{\text{DART}} = S^{\text{RT}} - S^{\text{DA}}
\]

Where C denotes congestion price, L denotes loss price, A/B denote any two nodes for which Up-to Congestion bidding is available, and DA, RT, DART denote the Day-Ahead, Real-Time, and Day-Ahead/Real-Time spread price.
The software compiles these values for each node pair for each hour for the past three years. All Up-to Congestion bids must have their source or sink node as one of eight specific nodes known as interface nodes, creating approximately 3250 node spread pairs. However in order to store the data most efficiently, we can store each node’s own price spreads separately, rather than redundantly combining each node pair, because we can rewrite the formula for DART spreads as follows:

\[
S_{DART} = S^{RT} - S^{DA}
\]

\[
= \left( C^{RT}_A + L^{RT}_A \right) - \left( C^{RT}_B + L^{RT}_B \right) - \left( C^{DA}_A + L^{DA}_A \right) + \left( C^{DA}_B + L^{DA}_B \right)
\]

\[
= S_A + S_B
\]

Thus we can dramatically reduce the size of the database by storing only intra-node DART spreads and recombining them in pairs to produce DART spreads.

The first database first is a large, static database that contains all six unique prices for each node on the entire PJM grid, which numbers approximately 10,000. This database is too large to be used at once, so if interested in data pertaining to nodes which do not allow Up-to Congestion bids, the data can be accessed, although more slowly.

The second distinct database is a small subset of the first, which is efficiently stored in order to be able to fit into the Random-Access-Memory (RAM) of a computer. The following smaller data sets are available for easy use in MATLAB:

- Day-Ahead node congestion and loss prices (DA prices)
- Real-Time node congestion and loss prices (RT prices)
- Day-Ahead/Real-Time intra-node spread (intra-node DART spreads)
- Day-Ahead internode congestion and loss spreads (DA spreads)
• Real-Time internode congestion and loss spreads (RT spreads)

• Day-Ahead/Real-Time internode congestion and loss spreads (DART spreads)

Below Figure 1 shows the data management GUI for importing data, inserting it into the existing database and monitoring other aspects. This GUI makes it possible to easily update data, and generate the necessary files for fast and easy analysis procedures.

Figure 1: Data Management GUI
The table on the right contains a list of all of the days included in the entire database, so the user can easily identify if the database is up-to-date, or potentially missing a day’s data. The text input allows the user to type in the name of the file containing the new set of data to insert, and a button to begin the process. Adding a day’s worth of data to the database takes approximately 90 seconds. Toggle buttons allow for the user to select a few options related to the GUI’s functionality. The software also supports the input of data when Day Lights Saving Time interrupts the usually constant number of hours in the day. Both databases are updated when new data is added using this GUI. In order to access data not in the smaller set, functions are needed to access nodes that do not support Up-to Congestion bids. Users can input the node identification number, and retrieve the relevant data.

Future iterations of this data management software will take advantage of PJM’s real time data protocols. Currently, the most recent data available for this software is delayed by one day, due to PJM’s data publishing method. However, HTML protocols exist for members of the PJM interconnection, which allows for data to be retrieved immediately after it becomes available, allowing this software to constantly update the database with the most current data. Along with the GUI shown above, a variety of other database management functions were developed.

**Map Research GUI**

In order to better interface with the large quantity of data, and to visualize how pricing works across the PJM grid, a map function was necessary to aid research. Because prices on the PJM grid are based on location, the relationship between the nodes is tightly linked to the
distance between nodes. Using a compiled database of longitude and latitude coordinates, the following map GUI shown in Figure 2 was constructed to show prices across the PJM grid.

Each dot on the map represents the location of a PJM node, though not all the PJM nodes’ location data is available. Only approximately 2.5% of the PJM nodes have publicly available latitude and longitude data. The color gradient, from green to yellow and red, represents the intra-node DART spread at that point for a given moment in time.

The controls in the GUI allow the user to input the Node Identification Number, or PnodeID, and highlight that point on the map, to show the recent pricing activity for the Day-Ahead and Real-Time prices. Additionally, it allows for the user to select a point on the map, and return the associated PnodeID. This facilitates research by allowing the user to visualize relationships in space between nodes.

Figure 2: Map GUI
The map easily shows how nodes are correlated highly based on their distance to each other, and how pricing can move across the grid with respect to weather and other variables. Below in Figure 3 is a scatter plot of the correlation coefficient between two nodes relative to the distance in degrees of arc length between them.

![Distance versus Correlation of nodes](image)

**Figure 3: Contemporaneous Correlation versus Distance**

This strong downward trend between correlation value and distance is readily shown visually using the interactive map GUI. Though nodes that are farther away tend to be less contemporaneously correlated, they become inversely correlated as the distance exceeds approximately 3 degrees of arc length.

Using this set of research tools developed to facilitate the research process, analysis of the underlying pricing data can begin. We will examine the impact DART spreads have on the following day’s DA spread, as well as the RT spread. If we can extract information from the relationship between DART spreads and the next day’s DA spread, we can improve bidding efficiency by more accurately estimating and bidding on the DA prices in the blind Dutch auction run by PJM.
Initial Descriptive Data

DART prices can be said to strongly resemble a normal Gaussian distribution. Below in Figure 4a shows a histogram of all DART spreads. A Gaussian curve fits with a 95.17% $R^2$ value, suggesting this description is more than accurate. However, substantial deviation occurs in the tails of the distribution, which do not fit the Gaussian assumption. Events that exceed three standard deviations from the mean occur with a frequency of 3.03%, versus an expected rate of 0.26% under the Gaussian assumption. Thus our assumption breaks down in the tail, as shown by the normal probability plot in Figure 4b. These markets show dramatic volatility.

![Figure 4a: Histogram of DART spreads](image)
Similarly, the average spread value for each node pair is again Gaussian, though with a slight shift towards average negative values, centered on -0.27, as shown in Figure 5a. Overall the system has a slight negative expected value. Figure 5b above shows a histogram of the relative probability of an intra-node DART spread being positive in any given hour. Though we might expect most spreads to have a 50% probability of overestimating or underestimating the real congestion price, there are almost no nodes that fall into this category. Instead, they tend to fall around the edges, at 30-45% or 65-70%, but very few fall around 50%. This suggests that two distinct sets of nodes may exist: those that tend to overestimate prices on a regular basis, and those that tend to underestimate prices. This effect may prove important in future research.

Figure 4b: Normal Probability Plot
Any efficient bidding strategy must account for this inherent difference in the behavior of each individual node. This finding is further substantiated using a correlation matrix of DA prices shown below in Figure 6, suggesting most nodes are either strongly correlated or inversely correlated with other nodes.
Principal Component Analysis

Given that contemporaneous correlations are high, it suggests the presence of strong principal components in the data. That is, we use singular value decomposition to decompose the pricing observations $X$:

$$X = W\Sigma V^T$$

Where $W$ is an eigenvector matrix of the covariance matrix of $XX^T$, $\Sigma$ denotes a diagonal matrix, and $V$ is the eigenvectors of $X^TX$. This allows us to decompose the data into orthogonal principal components. Figure 7 below shows the relative magnitude of the eigenvector decomposition, or the most significant principal components. We can see clearly that a small number of principal components are responsible for virtually all the variance present in the data.

![Figure 7: Principal Component Magnitudes](image-url)
Thus, it is important to include the potential influence of this small group of large principal components.

This project includes functionality for extensive research using principal component analysis techniques. Because this method breaks down data into orthogonal vectors, similar to analyzing signals using a Fourier transform, it is important to be able to identify specific principal components, and either remove or add specific components in order to examine the impact they have.

**DART spreads**

Given the previous background information, we can better approach the question of the impact of DART spreads on tomorrow’s DA prices. MATLAB’s built in functions give a powerful analytical tool for answering specific questions about the PJM pricing data. For this analysis, we will examine only the intra-node DART spreads. This eliminates the possibility of generating statically spurious relationships when comparing two spread pairs that share a common node.

If intra-node DART spreads for the previous day are high, that is, DA prices underestimated real time demand, then we might expect that to put upward pressure on tomorrows DA prices, to compensate for the previous day’s error. To verify if this hypothesis is correct, we can examine the cross correlation function of the intra-node DART and DA prices, shown below in Figure 8. This clearly shows a strong positive correlation between a DA price, and the intra-node DART spread which occurs 24 hours before for any given hour in the day. A peak occurs strongly at exactly 24 hours prior.
This analysis is flawed however, because each hour of the day has a different lag time between the RT price and the bidding time. For example, a 1:00AM contract has a 13-hour lag since bidding occurred. Midnight however, has a 36-hour lag. This difference information may further impact pricing.

![Sample Cross Correlation Function (XCF)](image)

**Figure 8: Cross Correlation DART vs. DA**

Given that the previous day’s intra-node DART spread clearly has some impact on the following day’s DA price level, we can look at the expected value of DA prices given yesterdays intra-node DART spread. Below in Figure 9 shows the expected DA price on the Y-axis, given the previous day’s intra-node DART spread, for all nodes across the grid. Given that yesterday’s
intra-node DART spread was high, we can clearly see that the expected price level of the following day’s DA price is higher, though prices tend to become much less predictable as the absolute value of the previous day’s intra-node DART spread increases. The opposite occurs when intra-node DART spreads are low.

Figure 9: DART versus DA Spreads

Though this trend is strong, it does not conclusively determine the relationship between the two variables. To better understand their connection, linear regressions can better determine the relationship.

If we run a linear regression between the previous day’s intra-node DART spread, and the next days DA price, we can see what percentage of the variance in the following day’s DA
price is attributable to the intra-node DART spread, using the $R^2$ statistic. Below is a histogram of $R^2$ regression values for a linear regression between the previous day’s intra-node DART spread on the DA price for each node.

Figure 10: R-squared Values

These coefficients tend to be relatively small and average only approximately 1.5%. However, given the large amount of data available, these figures are statistically significant. The majority of the variance in DA prices however, 98.5%, is due to other factors in the market, such and weather and other idiosyncratic market demand variations. These relationships are substantially more complex than a linear regression model, and thus this finding cannot be said to be conclusive.
Conclusions

Determining the impact of DA spreads based on DART spreads is a complex problem that requires analysis from many different fronts, all of which cannot be undertaken here. The tools now exist to fully answer these questions.

This project has allowed for Great Bay Energy employees to utilize tools for market analysis by constructing a database compatible with MATLAB, and allowing for the easy addition and management of the data. Additionally, various tools were created to address the specific needs of the data sets. Using these tools, we were able to come to a preliminary finding about the impact of DART spreads on DA prices, which can increase bidding efficiency. Additionally, the software built throughout this project will allow for the continued research of PJM market energy pricing, and may become of value to other Great Bay Energy employees who wish to use MATLAB as an analytical tool.
References:


Prescheduling Operations, PJM Manual 10, January 1, 2010
http://pjm.com/~media/documents/manuals/m10.ashx

http://pjm.com/~media/documents/manuals/m14a.ashx

Merchant Specific Transmission Requirements, PJM Manual 14E, July 5, 2005
http://pjm.com/~media/documents/manuals/m14e.ashx


http://www.pjm.com/pub/account/lmpgen/lmppost.html

https://edata.pjm.com/eContour/#
