Estimating Hourly Marginal Emission in Real Time for PJM Market Area Using a Machine Learning Approach

Caisheng Wang, Yang Wang, Carol J. Miller*
Department of Electrical and Computer Engineering
*: Department of Civil and Environmental Engineering
Wayne State University
Detroit, USA 48202
{cwang, fc4839, ab1421}@wayne.edu

Jeremy Lin
Interregional Planning Department
PJM Interconnection
Audubon, PA 19403, USA
Jeremy.Lin@pjm.com

***Abstract*—There has been no marginal emission information and/or marginal fuel mix data published by the regional transmission organizations (RTOs) or independent system operators (ISOs) in real-time. This paper presents a support vector machine (SVM) based method to estimate and predict hourly marginal emissions and marginal fuel mix in real-time in the PJM market area. Input to our SVM-based model includes a variety of publicly available data including the real-time locational marginal prices (LMPs), load demand, wind generation, historical marginal fuel data, and other information (such as day of the week and holidays). The results from the SVM are compared with real data from the years 2014 and 2015.**

***Index Terms*—Marginal Emissions, Marginal Units, Machine Learning, Support Vector Machine.**

I. Introduction

Climate change due to greenhouse gas (GHG) emissions has become one of the most challenging issues of modern society. To combat climate change, united efforts have been carried out worldwide to reduce GHG emissions, including the Kyoto Protocol, Copenhagen Summit, and the 2015 United Nations Climate Change Conference (UNCCC) that is to be held in Paris at the end of 2015. Although the Kyoto Protocol was not successfully implemented and there was not a workable and legal binding agreement after the Copenhagen Summit, society has been educated and awakened at an unprecedented level about the importance and seriousness of climate change issues. Moreover, many countries including the US and China are now taking realistic actions to reduce emissions production. It is also expected that "for the first time in over twenty years of United Nation negotiations, a binding and universal agreement on climate" will be achieved in the 2015 UNCCC [1].

Currently, the electricity sector is the largest source of GHG emissions because worldwide the majority of electricity is generated by fossil-fired power plants. In the U.S., electric power generation contributes about 40% of the total GHG, 64% of SO_2 emissions, 16% of NO_x emissions, and 68% of mercury air emissions as well as large shares of other pollutants such as small particulates in 2010 [2]. In August 2015, the U.S. issued its first ever national standards to regulate carbon emissions from electric power plants through the Clean Power Plan (CPP) [3]. The objective of the CPP is to achieve reductions in emissions of CO_2, SO_2, and NO_x (32% 90% and 72%, respectively, all compared to the levels of 2005 [3]).

Electricity is finally consumed by end users in homes, businesses, and factories. Therefore, the behavior of end users can play an important role in reducing emissions, which has been largely overlooked so far. If the GHG emissions from electricity production can be attributed to the electricity consumption of the various end users, it will be very useful for understanding the influence of individual behaviors and making demand management more targeted and highly efficient. Studies have shown that electricity consumers are willing to change their behavior if additional information is available [4]. In addition to demand management, the impacts on emissions due to other smart grid applications such as large scale energy storage systems and electric vehicles have also generated significant research interest [5], [6]. It is then critical to correctly account for the emissions related to electricity generation, delivery and consumption.

Although the emissions can be monitored and measured directly at the power plants, information about the real-time marginal emissions due to marginal electricity generation are not available to the public. The classical electric energy system is unidirectional and top-down oriented. For example, in the U.S., power generation and consumption are coordinated through regional energy markets managed by regional transmission organizations (RTOs) or independent system operators (ISOs) [7]. RTOs and ISOs own all their system information and the majority of that information is not released to the public due to security and market confidentiality reasons even though the RTOs/ISOs have made significant progresses in making more data publicly available. Locational Marginal Prices (LMPs) have been posted publicly at commercial pricing nodes (CPNs) every 5 mins in markets operated by RTOs/ISOs. However, up until now, there are no locational

The work in this paper is supported by the National Science Foundation of USA under Grant ECCS-1508910. The views presented in this paper do not necessarily reflect those of the authors' affiliations.

marginal emission indicators to help electricity consumers manage their electricity usage for emission reductions. The objective of the optimal power flow (OPF) models currently used in RTOs/ISOs is to minimize generation costs, and therefore the LMP value [8] is not a direct index of emission impacts. Even though emission caps can be enforced theoretically in OPF models of ISOs and RTOs [9], the fact that the LMP is designed for cost reduction only will not change for the foreseeable future. As a result, although annual or monthly air emissions from electrical generation are well documented for all regions due to reporting requirements by the US Environmental Protection Agency (EPA) and Energy Information Administration (EIA), there is no information source for accurate real-time emissions data [10]. This lack of real-time emissions information hinders the ability to make control decisions based on the amount of emissions that would be generated at any time and makes it difficult for electricity end users to understand GHG emissions and to actively participate in the process of emission reductions. Therefore, it would be beneficial to have a tool that provides estimated marginal emissions in real-time or near real-time to guide the consumers' behavior.

The problem of estimating marginal emissions in power systems has received significant research attention. A bidding strategy for electricity generating companies that optimizes both cost and emissions was modeled in [11]. It was demonstrated that emissions and costs can be optimized simultaneously with dual benefits. Long-range marginal emission factors for CO_2 were modeled in the British electricity system based on expected changes to fuel mixes and power plants [12]. CO_2, SO_2, and NO_x were all reduced by a LEEM-based spatial shift in load more strongly than a strategy based on minimizing LMPs alone. LMPs were also used to determine the effect of electricity trading between Quebec with New York and New England on CO_2 emissions [13]. Using probability distribution functions of fuel costs and generation unit heat rates, we have developed an LMP-Emissions Estimation Method (LEEM) to map real time and day-ahead LMPs to marginal generator types, i.e. fuel oil, coal, natural gas, or nuclear/renewable/hydro [10], [14]. The LEEM model was used in [15] to drive an apportionment of electricity loads among several locations to minimize emissions. The work presented in this paper is an extension to the work we have done on the LEEM. Based on the publicly available data for the PJM area, a machine learning approach is proposed to estimate and predict hourly marginal emissions at the regional level. The results obtained are verified and compared against the real marginal fuel mix information for the region.

The remainder of the paper is organized as follows: The concept of marginal emissions and publicly available data from PJM market are presented in detail in Section II. A support vector machine (SVM)-based method is introduced in Section III to link the LMP, load demand and wind generation data in the PJM area and other information (such as weekdays, holidays, etc.) to the hourly marginal fuel mix in the region. Preliminary results regarding to marginal emission prediction and marginal fuels prediction are shown in Section IV. The results are compared with the actual marginal fuel mix data in the PJM region in the same section, followed by the conclusions drawn in Section V.

II. Marginal Emission and Publicly Available Data Sources

A. Marginal Units and Marginal Emission

Marginal units in a real-time electricity market are the units which respond to load changes and set the locational marginal prices in each five minute interval. Marginal units are identified by running security-constrained economic dispatch (SCED) in real-time or day-ahead electricity markets. In this paper, only the real-time market is considered. The objective of SCED (also called optimal power flow) is to minimize the total generation cost of the whole system while observing the transmission constraints and generator capacity limits. Therefore, the system congestion conditions and the generator cost models (or bidding models) have significant impact on the determination of marginal units.

In this paper, the marginal emissions are defined as the weighted emission rates of the marginal units which may consume different fuels, given in the following equation:

$$\gamma^* = \sum_{i=1}^{n} \gamma_i k_i \tag{1}$$

where γ^* is the marginal emission; $\mathrm{\gamma_i}$ is the *i*th marginal generation/fuel type and $\mathrm{k_i}$ is the corresponding percentage of the fuel type over the period under study (i.e. one hour in this paper) and *n* is the total number of fuel types in a given hour. The percentage $\mathrm{k_i}$ of each fuel type in each hour is calculated based on the number of five minute intervals that the fuel type is marginal or jointly marginal; and $\sum_{i=1}^{n} k_i = 1$.

B. Publicly Available Data

The PJM market publishes the hourly marginal fuel types at the system level, but with a delay of two months [16], [17]. Table I shows an example of the marginal fuel mix information published for the first two hours of September 01, 2015. At hour 0 (i.e. from midnight to 1 am) on that day, the marginal units in the whole PJM area consist of 75% coal-fired generators and 25% natural gas-fired generators. However, there is no real-time information available regarding marginal units and associated marginal fuel types. Therefore, there is a need to predict the fuel types and emissions in real-time based on the historical information. The goal is to help customers manage their electricity consumptions in reducing emission and help carry out an accurate life-cycle analysis of electricity generation that is important for various products and processes including GHG inventories of different entities and regions.

Hourly (as well as 5-min) LMP data, hourly load demand data and wind generation in the PJM region are publicly available [16]. As aforementioned, marginal units are dependent on the current system operating state, generator cost models, congestions, wind generation and load levels. In line with the publicly available data from PJM, the data of LMP, load and wind generation are used as the inputs in the proposed forecasting model that is discussed in the following section. LMPs can reflect the system congestion conditions. They, together with the wind and load data, can implicitly reflect the current system state. The output of the forecasting model is weighted emission rate associated with the fuel types of marginal units, which is published with a two-month delay [17]. The time granularity in the forecasting model is one hour. The detailed information of the data for the years 2014 and 2015, utilized in our model is listed in Table II.

TABLE I.
FORMAT OF FUEL TYPES OF MARGINAL UNITS AT PJM

DATE/HOUR	TIME ZONE	FUEL TYPE	PERCENT MARGINAL
01SEP2015:00:00:00	EDT	Coal	0.75
01SEP2015:00:00:00	EDT	Natural Gas	0.25
01SEP2015:01:00:00	EDT	Coal	0.6667
01SEP2015:01:00:00	EDT	Natural Gas	0.3333
…	…	…	…

TABLE II.
KEY ATTRIBUTES OF PUBLICLY AVAILABLE DATA FROM PJM

	Load	Wind	LMP	Fuel
Hourly Average	'RTO','RFC', 'MIDATL', 'WEST', 'SOUTH', 'PS','PE', 'PL','BC', 'GPU','PEP', 'CNCT', 'RECO','AP', 'CE’,'AEP', 'DAY', 'DUQ', 'DOM', 'ATSI', 'PLCO', 'UGI','JC', 'ME','PN', 'AE','DPL', 'OE', 'PAPWR', 'DEOK'	'RTO', 'RFC', 'MIDATL', 'WEST', 'SOUTH'	'PJM', 'AECO', 'AEP', 'APS', 'ATSI', 'BGE', 'COMED', 'DAY', 'DEOK', 'DOM', 'DPL', 'DUQ', 'EKPC', 'JCPL', 'METED', 'PECO', 'PENELEC', 'PEPCO', 'PPL', 'PSEG', 'RECO'	'Coal’, ‘Natural Gas’, ‘Light Oil’, ‘Waste Coal’, ‘Municipal Waste’, ‘Land Fill Gas', 'Wind', 'Miscellaneous', 'Kerosene', 'Heavy Oil', 'Diesel', 'Demand Response', 'Min Gen/Dispatch Reset', 'Solar', 'Uranium', 'Virtual Sale at MISO', 'Virtual Sale at NY', 'Biomass', 'Missing Data'
Summary	30 regions and load zones with some overlaps	5 regions	20 price zones plus RTO price zone	19 fuel types

As shown in Table II, there are 30 load regions, 5 regions with wind generation data reported, 21 LMP zones and a total of 19 fuel types in the years of 2014 and 2015. Some of the fuel types rarely appear as the marginal units. The marginal units are further grouped into five fuel types: coal, natural gas, petroleum/oil, wind and miscellaneous, as summarized in Table III. The corresponding emission rates with respect to CO_2, SO_2 and NO_x are listed in Table IV [15].

TABLE III.
GROUPED FUEL TYPES: AN EXAMPLE FOR A TYPICAL MONTH

Period	Coal	Natural Gas	Petroleum	Wind	Miscellaneous
2014	0.4895	0.3634	0.0929	0.0457	0.0085
2015(Jan-Aug)	0.5024	0.3902	0.0587	0.0404	0.0083

TABLE IV.
EMISSION RATES OF DIFFERENT FUEL TYPES

Fuel Types	Coal	Natural Gas	Petroleum	Wind
CO2(lbs/MBtu)	210.97	101.16	134.62	0
SO2(lbs/MBtu)	1.2195	0.0089	0.9662	0
NOx(lbs/MBtu)	0.5629	0.1212	0.3221	0

Based on the information given in Tables I, III and IV, the weighted marginal emissions can be readily calculated using (1). Since the share of “Miscellaneous” fuel types is very small (less than 1%), it is neglected in the marginal emission calculation.

III. SUPPORT VECTOR MACHINE METHOD AND PROBLEM FORMULATION

The purpose of the paper is to estimate and predict hourly marginal emissions based on the historical data of LMP, load, wind and fuel mix that are publicly available. The support vector machine (SVM) [18]-[20] is used to fulfill the task. A brief introduction of SVM is given hereby.

For given samples $G = \{(x_i, y_i)\}_i^N$, the general idea of support vector machines (SVM) is that an original feature space, i.e., x spaces, can always be mapped to some higher-dimensional feature space, i.e., $\phi(x)$ spaces, so that the linear regression can be used to catch the relationship between $\phi(x)$ and y.

$$f(x) = \sum_{i=1}^{D} w_i\, \phi_i(x) + b \tag{2}$$

where b and w_i are coefficients and can be estimated by minimizing

$$\frac{1}{N}\sum_{i=1}^{N} |f(x_i) - y_i|_\varepsilon + \lambda \|w\|^2 \tag{3}$$

where λ is a constant and the cost function is defined by

$$|f(x_i) - y_i|_\varepsilon = \begin{cases} |f(x) - y| - \varepsilon & for\ |f(x_i) - y_i| \geq \varepsilon \\ 0 & otherwise \end{cases} \tag{4}$$

where ε is the width of the insensitive tube.

The dual problem of (3) is

$$\min_{\alpha_i,\alpha_i^*} \frac{1}{2}(\alpha_i + \alpha_i^*)(\alpha_i - \alpha_i^*)K(x_i, x_j) + \sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \varepsilon \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \tag{5}$$

Subject to

$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \tag{6}$$

$$0 \leq \alpha_i, \alpha_i^* \leq C, \qquad i = 1, \dots, l \tag{7}$$

where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is called Kernel. Often used Kernels include

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial of power p: $K(x_i, x_j) = (1 + x_i^T x_j)^p$
- Gaussian (radial-basis function): $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$

There are only a few of non-zero coefficients in α_i and α_i^*, the data points associated with them are called the support vectors. Referring to [21], the input variables of the proposed SVM model are designed as the ones given in Table V. It should be noted that each hour has a SVM model. In other words, there are a total of 24 SVM models corresponding to the 24 hours in a day. Meanwhile, Gaussian kernel function is adopted in our models.

TABLE V.
INPUT/OUTPUT VARIABLES IN SVM

Input	**Mark**	**Description**
Load	L	30 loads in different PJM load zones
Wind	W	5 wind generation points
LMP	P	21 LMPs at different price zones
Day of the week	$W_{1,...,7}$	7 binary values for each day in a week
Month	$M_{1,...,12}$	12 binary values for each month in a year
Holiday	$H_{1,...,4}$	4 binary values for indicating holiday
Total	[M,W, H, L,W,P]	
Output	Y	Percentages of the 5 fuel types identified in Table III

IV. Preliminary Results

The two years of data for 2014 and 2015 are used to train and test the proposed SVM model. This set of data is divided into two parts: the training data, i.e., a total of 540 points with respect to the days between Jan. 1, 2014 and Jun. 24, 2015 and the testing data, i.e., a total of 60 points between Jun. 25, 2015 and Aug. 23, 2015. The training data is further divided into ten folds and a cross validation technique is used to select the parameters of the SVM models.

The mean absolute percentage error (MAPE), used to evaluate the performance of the proposed forecasting model, is defined as:

$$\mathrm{MAPE} = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{\hat{y}_i}\right|\right)\cdot 100\,\% \qquad (8)$$

where n is the number of test data; y_i and $\hat{y}_i$ are the real values and the predicted results, respectively.

A. Marginal Emission Prediction

The predicted marginal emissions (CO_2) are shown in Fig. 1. For comparison, the real published values are also plotted in the same figure. The MAPEs at different hours are calculated and listed in Table VI.

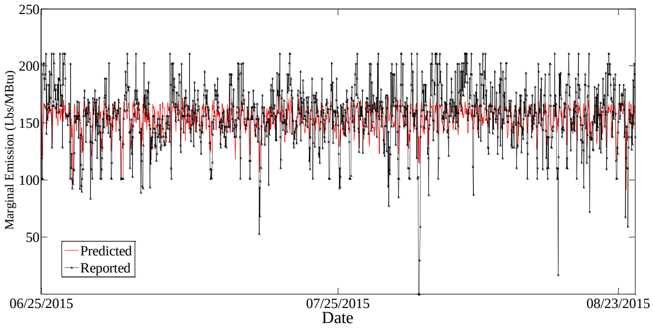

Fig. 1. Predicted and Reported Marginal Emissions (CO_2) in PJM.

It can be observed that the 4th hour SVM model has the biggest MAPE error, i.e. 19.75%, and the smallest error, i.e. 7.04%, appears in the 18th hour SVM model. The study results validate the effectiveness of the proposed SVM model, with an acceptable performance for most hours of the day.

TABLE VI.
MEAN ABSOLUTE PERCENTAGE ERRORS (HOURLY)

Hour	**1**	**2**	**3**	**4**	**5**	**6**
MAPE	14.03	15.04	18.85	19.75	16.34	15.85
Hour	**7**	**8**	**9**	**10**	**11**	**12**
MAPE	14.13	10.51	9.81	11.76	10.71	11.21
Hour	**13**	**14**	**15**	**16**	**17**	**18**
MAPE	10.00	10.31	10.02	8.67	9.85	7.04
Hour	**19**	**20**	**21**	**22**	**23**	**24**
MAPE	8.34	9.04	9.79	9.4	8.53	14.6

B. Marginal Fuel Mix Prediction

Besides the estimation and prediction of the overall marginal emissions, the percentages of different marginal fuels can be estimated and predicted as well. The same SVM models used in the overall marginal emission prediction are used in this preliminary study for marginal fuel mix prediction. The only difference is that the outputs are changed to the percentages of different fuels, i.e. coal, natural gas, petroleum and wind. The results are shown in Figs. 2-4. Although the predicted fuel mix closely follows the actual reported data, there are relatively larger errors when the SVM models are used to predict the marginal fuel share percentages. This might be due to the fact that the change in the share of an individual marginal fuel can be more volatile since the load demand, wind generation and system constraints are highly stochastic.

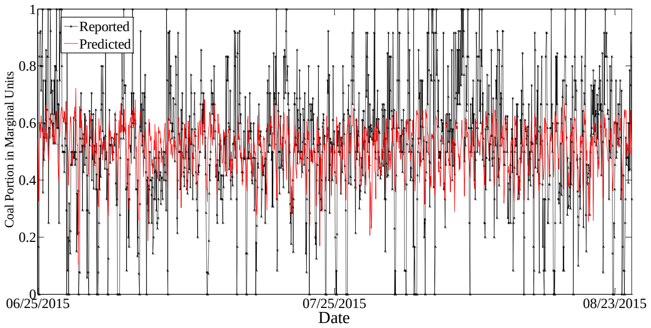

Fig. 2. Share of Coal as a Marginal Fuel.

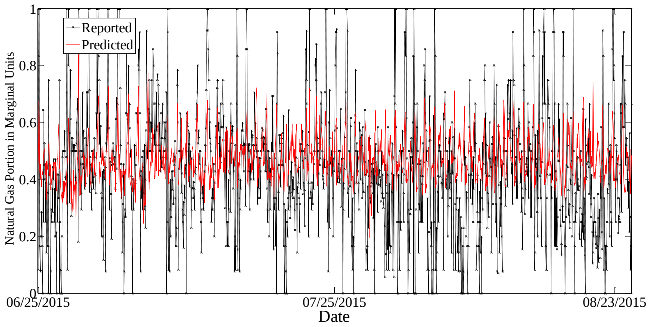

Fig. 3. Share of Natural Gas as a Marginal Fuel.

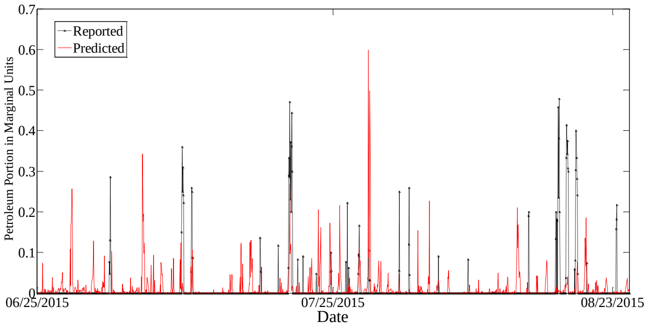

Fig. 4. Share of Petroleum as a Marginal Fuel.

As shown in Figs. 1-4, even though there is room for further improvement, the preliminary results provide optimism in the ability of the method to provide customers useful information about the marginal emissions associated with their electricity consumption. To our best knowledge, the work in this paper is the first of its kind in predicting real-time marginal emissions due to electricity generation in a market setting. As a future task, the SVM models can be updated to include more publicly available information, such as the system congestions (i.e. binding constraints), shadow prices, weather conditions and other useful historical data. Other machine learning methods can also be used in place of SVM for this purpose in the future.

V. Conclusion

This paper introduced an SVM-based model for the prediction of real-time marginal emissions and marginal fuel mix for the PJM market area. Case studies using the data from 2014 and 2015 have been carried out to validate the results from the proposed method. The preliminary results show that the proposed approach is effective in predicting the overall hourly marginal emissions in real-time. Nevertheless, the results also show that there is room for further improvement in marginal fuel mix forecast. For future work, the proposed SVM model can be enhanced to include additional publicly available information to further improve the prediction performance. Other machine learning algorithms such as deep learning can also be explored for this task. Ultimately, the availability of real-time marginal emission information which will enable energy consumers to develop more environmentally sensitive energy practices.

References

[1] *COP - What's it all about?* Online: http://www.cop21paris.org/about/cop21, retrieved on 11/4/15.

[2] C. Van Atten, A. Saha, and L. Reynolds, "Benchmarking Air Emissions of the 100 Largest Electric Power Producers in the United States," 2012 Report, Online: http://www.nrdc.org/air/pollution/benchmarking/files/benchmarking-2012.pdf.

[3] *FACT SHEET: Overview of the Clean Power Plan, Cutting Carbon Pollution from Power Plants*, Online: http://www2.epa.gov/cleanpowerplan/fact-sheet-overview-clean-power-plan, retrieved on 11/4/15.

[4] C. Cédric, "Smart grids: another step towards competition, energy security and climate change objectives," *Energy Policy*, vol. 39, no. 9, pp.5399–5408, 2011.

[5] M. Tamayao, J. Michalek, C. Hendrickson, I. L. Azevedo, "Regional variability and uncertainty of electric vehicle life cycle CO2 emissions across the United States," *Environmental Science & Technology*, vol. 49, no. 14, pp 8844–8855, 2015.

[6] E. Hittinger, I. L. Azevedo,, "Bulk Energy Storage Increases US Electricity System Emissions," *Environmental Science & Technology*, vol. 49, no. 5, pp. 3203-3210, 2015.

[7] Online: https://www.ferc.gov/market-oversight/mkt-electric/overview.asp.

[8] S. Stoft, *Power system economics: designing markets for electricity*, IEEE press. Wiley-Interscience; 2002.

[9] Lower emissions cap for Regional Greenhouse Gas Initiative takes effect in 2014, Online: http://www.eia.gov/todayinenergy/detail.cfm?id=14851.

[10] M. M. Rogers, Y. Wang, C. Wang, S. P. McElmurry, C. J. Miller, "Evaluation of a rapid LMP-based approach for calculating marginal unit emissions," *Appl. Energy*, vol. 111, pp. 812–820, 2013.

[11] V. Vahidinasab, S. Jadid, "Multiobjective Environmental/Techno-Economic Approach for Strategic Bidding in Energy Markets," *Applied Energy*, vol.86, pp. 496-504, 2009.

[12] A. D. Hawkes, "Long-Run Marginal CO2 Emissions Factors in National Electricity Systems," *Applied Energy*, vol. 125, pp. 197-205, 2014.

[13] M. B. Amor, P.O. Pineau, C. Gaudreault, R. Samson, "Electricity trade and GHG emissions: Assessment of Quebec's hydropower in the Northeastern American market (2006–2008)," *Energy Policy*, vol. 39, pp.1711-1721, 2011.

[14] C. Wang, Carol J. Miller, Timothy H Carter, Shawn P. McElmurry, Michelle Rogers, Stephen S. Miller, and Ian A. Hutt, "Linking Load Demands to Power Generation Pollutant Emissions Based on Locational Marginal Prices," *IEEE PES Transmission &Distribution Conference and Exposition*, 2012.

[15] Y. Wang, C. Wang, C. J. Miller, S. P. McElmurry, S. S. Miller, and M. M. Rogers, "Locational Marginal Emissions: Analysis of pollutant emission reduction through spatial management of load distribution," *Applied Energy*, vol. 119, no. 15, pp. 141–150, 2014.

[16] PJM marginal fuel type data, Online: http://www.pjm.com/markets-and-operations/energy.aspx.

[17] Marginal Fuel Posting, Online: http://www.monitoringanalytics.com/data/marginal_fuel.shtml

[18] V. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.

[19] S. Mukherjee, E. Osuna, F. Girosi F. "Nonlinear prediction of chaotic time series using support vector machines," *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, Amelia Island, FL, pp. 511–20, September 1997.

[20] K. R. Muller, A. Smola, G. Ratsch, B. Scholkopf, J. Kohlmorgen, V. Vapnik. *Predicting time series with support vector machines. Advances in kernel methods—support vector learning*, Cambridge, MA: MIT Press, pp. 243–254, 1999.

[21] E. Ceperic, V. Ceperic, A. Baric, "A strategy for short term load forecasting by support vector regression machines," *IEEE Trans Power System*, vol. 28, no. 4, pp. 4356–4364, 2013.