Energy Grid Forecasting and Anomaly Detection

Cody Hill
University of Colorado Boulder
Boulder, USA
cody.hill-1@colorado.edu

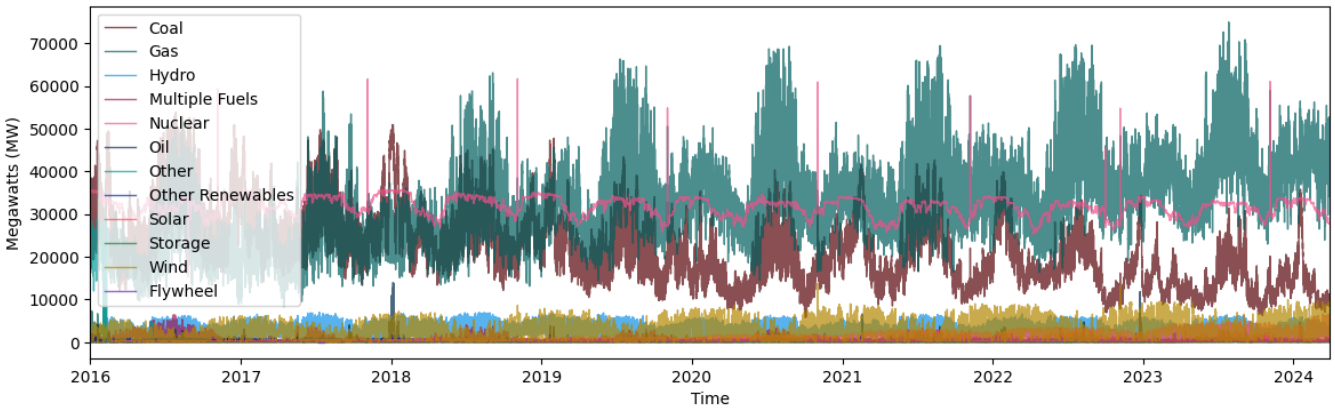

Figure 1: Energy Generated in Megawatts (MW) by Fuel Type.

ABSTRACT

Energy forecasting is a vital part of today's energy markets and operations. It is used to set the cost of electricity, determine when power generation should be increased, and decide when power needs to be directed to or from other regions of the electrical grid. Electrical energy demand continues to rise, especially with the recent widespread adoption of electric vehicles. As a result, the electrical grid will need to become more robust and adaptive as this increase in demand comes in tandem with a transition away from fossil fuels and into a more diversified energy market and with it, more complex cycles. These circumstances make accurate forecasting periodicity of electrical loads and generation that much more important, and it is necessary to research the differences in forecasting applications and relevance in various short-term, medium-term, and long-term settings.

Many of the statistical time series forecasting techniques of the past are still in use today, each with their own strengths and weaknesses, often trading compute efficiency with lesser accuracy in long-term forecast horizons. More modern machine learning models and neural network based architectures tend to generalize through forecasting horizons better, but at the cost of compute resources. Though because of the potential for complexity of energy load patterns, both categories see benefits in task-specific tuning. By evaluating each forecasting technique's accuracy across various context windows and forecasting horizons, this project presents a thorough comparison on which techniques are most relevant in the task-specific energy grid context and forecast horizons.

Data Mining Project '24, April 2024, Boulder, CO
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Mathematics of computing** → **Time series analysis**; • **Computing methodologies** → **Supervised learning by regression**; *Classification and regression trees*; *Support vector machines*; *Neural networks*; • **Applied computing** → **Forecasting**.

KEYWORDS

time series forecasting, anomaly detection, energy load, renewables, energy grid

ACM Reference Format:
Cody Hill. 2024. Energy Grid Forecasting and Anomaly Detection. In *Proceedings of Data Mining Project (Data Mining Project '24).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 AUTHOR'S NOTE

This document is to be used as a project proposal for the beginning planning stages of the project as well as a living-document as the project progresses. Up until project completion when a new version will be created to fill in the results.

2 INTRODUCTION

Energy forecasting is an important task to not only determine the expected load on the energy grid and the requirements of energy providers to meet that demand by either ramping up production or diverting sources altogether. Identifying hourly, daily, and seasonal energy demands accurately has huge implications on not only the infrastructure of our energy grid but also a complex energy marketplace where prices dynamically shift based on generation supply, customer demand, and financial spot and derivative contracts which are based on energy forecasting.[x] In the past this predictive role was largely upheld with statistical point-estimate models, but with

Table 1: Defining prediction forecast horizon intervals. The forecast horizon determines how far into the future a forecast model predicts.

No.	Forecast Horizon Intervals	Duration
1	Very Short-Term	$X < 1$ Hour
2	Short-Term	1 Hour $\leq X <$ 1 Week
3	Medium-Term	1 Week $< X \leq 1$ Year
4	Long-Term	$X > 1$ Year

the growth of our predictive models along with the diversification of the energy grid more modern probabilistic interval forecasting techniques have begun to be used.[x] This diversification of the energy grid not only comes from a dismantling of the monopolistic qualities of the energy market in the 1990's,[x] but emerging cultural desires and improved technology in renewable energies are driving a need for a better electrical grid.[x]

Electricity generation has always been mostly an on-demand industry and largely continues to be, where the energy we generate needs to be transmitted in used within a short period of time because we lack efficient systems for energy storage at scale.[x] This fact presents additional complications when the majority of renewable energies only work in certain conditions, making its effective use more reliant on forecasting than traditional energy generation technologies. However, this means with accurate forecasting, we can reduce reliance on fossil-fuel energies by more efficiently filling in gaps of energy demands with renewables.

3 RELATED WORK

Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing models tend to be useful in short-term forecasts, but can experience limitations modeling complex periodocity. Newer, but certainly not state-of-the-art, machine learning techniques such as Gradient Boosting Machines or XGBoost and Long Short-Term Memory (LSTM) network models tend to generalize better and excel at longer-term forecasts.

4 METHOD AND PROPOSED WORK

Models stemming from different statistical and machine-learning families will be trained to perform electricity load forecasting on the forecast horizons listed in Table 1. Each model's amount of compute required for training and evaluation scores will be compared on different forecast horizons. See evaluation section for more details on the specific metrics used.

4.1 Data

Data was collected by the Pennsylvania-New Jersey-Maryland Interconnection (PJM). PJM is a regional electrical transmission organization which coordinates electricity transmission in all or parts of Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia and the District of Columbia.[x] A map of this can bee seen in Figure 2.

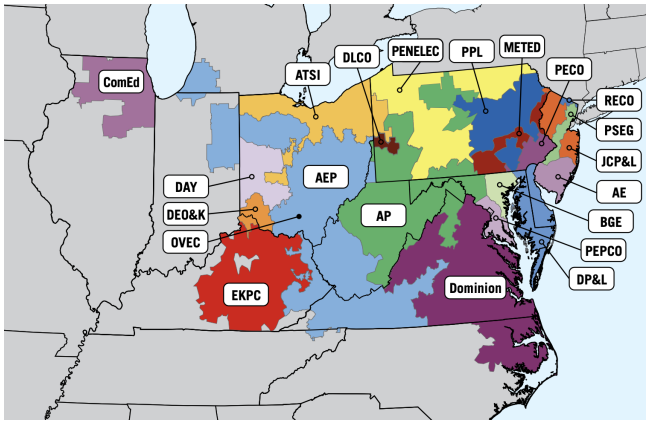

Figure 2: Regional Transmission Zones Labeled by Distributor IDs.[x]

Two sources of data from PJM are being utilized here, hourly load in megawatt-hours verified by the individual electric distribution companies and hourly electrical generation aggregated by fuel type, also in megawatts.[x] The hourly load data is flagged with geographical location data which will be leveraged to include regional weather data as a covariate in the models. An overview of the datasets can be seen in Table 2.

Table 2: Dataset size and date ranges.

Dataset	Size (Rows)	Date Range
Hourly Load	5,446,243	1993/01/01 - 2024/03/27
Hourly Generation	783,663	2016/01/01 - 2024/03/28
Avg Hourly Weather	TBD	TBD

4.2 Tools and Techniques

First these three datasets will be datetime aligned and analyzed for an outliers. The different generation sources and load data will then be analyzed and plotted for historical trends and to identify periodicity (cycles) at different context window intervals which will help set expected baselines for the forecast horizons. Once the data has been properly aggregating and aligned the data will then be split into a training, validation, and test split [METHOD TBD].

Models of varying complexity and methodology will be used to better represent the broad spectrum of techniques used today, and to attempt to capture a good representation of which models work best for the different forecast horizons.

Proposed models:

Baseline

- Naive Forecast Model

Statistical

- Autoregressive Integrated Moving Average (ARIMA)
- Vector Autoregression (VAR)

Machine Learning

- Support Vector Machine (SVM)

- XGBoost

Deep Learning

- Long Short-Term Memory (LSTM) Networks
- Transformer Networks

Pre-trained

- Time Series to Vector (TS2Vec)

Some models may require data transformation or normalization to perform effectively. These changes will be listed here.

5 EVALUATION

Symmetric Mean Absolute Percentage Error (sMAPE) will be the univariate evaluation metric used to compare all models. This will ensure comparisons of error across different models and forecasting horizons are normalized.

$$\text{sMAPE} = \frac{200}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} \tag{1}$$

where y_i and $\hat{y}_i$ are the actual values and forecasted values respectively, and n is the number of points in the set.

One possible limitation of sMAPE is if a predicted value or actual value equals zero, the metric balloons to the maximum error value, and if both equals zero, the metric is undefined. Chosen evaluation metrics often have strengths and weaknesses, and it is yet unclear if this limitation will occur here. A consideration will be made on a different choice if it does, especially since power generation by different fuel types may have a periodicity that includes zero power generation.

6 DISCUSSION

Discussions and analysis will be here when the project has been completed.

6.1 Project Timeline

6.2 Potential Challenges

A few challenges may arise impeding progress or causing the project to fall behind schedule are expected to be the following. Generally, properly fitting the data to each proposed forecast model within the timeline maybe pose a challenge. Specifically, the neural-network architectures can take considerable time to build from the ground up without intimate knowledge and experience of their specific architectures. Also, completely fleshing out the three goals of the project on schedule may be cause for concern and need reevaluation – anomaly detection, hourly load forecasting, and hourly energy generation forecasting.

6.3 Alternate Approaches

If the project is running behind the timeline a few different approaches can be considered to make the target deadlines.

(1) The number of models can be reduced but leaving at least one remaining in each model family type.
(2) In the case of some of the neural-network based architectures and machine-learning approaches, complexity can be reduced by utilizing more out-of-the-box approaches/models.
(3) Instead of forecasting both hourly load usage and generation by fuel type, the scope of the project can be reduced to focus on just one of these domains.

Table 3: A general framework onto which the project will progress. This timeline is subject to change.

Timeline	Agenda Item	Progress	Done
Days 1 - 5	Data Mining	**X**	
	Hourly Load Data	–	**X**
	Hourly Generation Data	–	**X**
	Weather Data	**X**	
	Cleaning/Preprocessing	**X**	
	Datetime Alignment	**X**	
	Aggregate by Zones	**X**	
	Outlier Inspection		
	Corrupted Data Removal		
	Exploratory Data Analysis		
	Train/Val/Test Split		
Days 6-8	Anomaly Detection		
	Model Setup		
	Training		
	Val. Optimizations		
	Build Eval. Functions		
Days 9-10	Test Results		
	Discussion		
	Conclusion		

7 CONCLUSION

By analyzing and modeling energy time series data we can efficiently and accurately forecast into the future different interval sizes, or forecast horizons, which correspond to real-world decisions. Modern techniques in time series forecasting have reduced the errors in all forecast horizon intervals and continue to generalize those prediction abilities across many domains. This type of research and technology is necessary for the uninterrupted energy generation and distribution due to the ever-increasing demands on our energy grids.

7.1 Key Findings

To be completed upon project completion.

7.2 Future Work

To be completed upon project completion.

8 APPENDICES

PLACEHOLDER

ACKNOWLEDGMENTS

Sqyd, for always brewing the morning coffee – I know it's 8 cups of water, now.

Luna, for breaking up the hyper-focused monotony with mind-clearing afternoon walks.

A RESEARCH METHODS

B ONLINE RESOURCES