Group 9 – Project Proposal

CHEMENG 4H03: Big Data Methods and Modeling in Chemical and Materials Engineering

Professor Alexandre D'Souza

6 March 2025

Giulia Morris-Cefis – morrig13 – 400376054

Kurt Snell – snellk1 – 400254458

Madeline Wighardt – wighardm – 400253663

Data Source

The dataset for this project was sourced from Kaggle [1] and compiles four years of hourly electrical consumption, generation, pricing, and weather data for cities in Spain. The data originates from multiple sources: consumption and generation data were retrieved from the European Network of Transmission System Operators for Electricity (ENTSOE), settlement prices were obtained from Red Eléctrica de España (REE), and weather data was collected through the Open Weather API for Spain's five largest cities, Valencia, Madrid, Bilbao, Barcelona, and Seville. The dataset was initially compiled for research into energy market forecasting and its role in transitioning to a renewable-based electrical infrastructure, research paper detailing this dataset and its applications was titled 'Tackling Climate Change with Machine Learning' by Rolnick et al [2].

Dataset

This dataset provides a comprehensive view of Spain's electricity market, including real-time and forecasted consumption, generation by energy source, market prices, and corresponding weather conditions. The availability of hourly data makes it valuable for predictive modeling and benchmarking against existing industry forecasts. Through analysis of this dataset, we aim to explore potential correlations between weather and energy variables, predict energy consumption based on weather conditions, and/or discover temporal trends in energy consumption.

Preliminary Ideas

Independently, the latent variable models and other models can be created from the weather dataset. Preliminary ideas include:

- Predicting wind speed from temperature, pressure, and humidity
- Determining historical weather conditions across time
- Comparing weather across localities

The electricity use dataset can also be analyzed independently:

- Determining energy generation from different sources as a function of weather conditions
- Determining energy pricing as a function of generation source
- Tracking historical energy consumption over time periods
- Accuracy of forecasted load over time

Taking the two datasets together, more complex analyses can be made:

- Accuracy of forecasted load according to weather conditions
- Determining energy consumption as a function of weather conditions
- Determining energy price as a function of weather conditions

To achieve the data analysis, we first must combine the two datasets. The datasets are indexed by date and time, so combination should be straightforward, although the location data must be considered. To reduce the dimensions of the dataset, find correlations between variables, and identify the most significant variables, an LVM may be performed. Alternatively, an ANN could be constructed to predict energy consumption amount, type, or price based on weather conditions, time and/or location.

References

[1] Kolasniwash, “Hourly energy demand generation and weather,” *Kaggle*, 2020. [Online]. Available: https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather

[2] D. Rolnick, P. Donti, L. Kaack, K. Kochanski, and A. Lacoste, “Tackling Climate Change with Machine Learning,” *arXiv*, Nov. 2019. [Online]. Available: https://arxiv.org/abs/1906.05433