Group 9 – Project Checkpoint

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

Professor Alexandre D'Souza

30 March 2025

Giulia Morris-Cefis – morrig13 – 400376054

Kurt Snell – snellk1 – 400254458

Madeline Wighardt – wighardm – 400253663

Preliminary Analysis of Dataset

The dataset was sourced from Kaggle [1] and was initially compiled for research into energy market forecasting and its role in transitioning to a renewable-based electrical infrastructure. A research paper detailing this dataset and its applications was titled 'Tackling Climate Change with Machine Learning' by Rolnick et al [2]. The data obtained consists of two separate datasets that compile four years' worth of hourly data from five cities in Spain on electrical consumption, generation, and pricing, and the second on various weather data. 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 ENTSOE represents 40 electricity transmission systems from 36 countries across Europe, they are committed to ensuring optimal functioning and development of the European electricity markets [3]. Their 'transparency platform' is a public access central collection and publication of electricity transportation data, consumption data, and information for the pan-European market [3]. The REE publish economic data to public domain to ensure reliable exploitation of the Spanish Power System in real time [4]. Data from Open Weather API is collected from different weather stations and is available for purchase, the data was purchased for the purposes of the paper by Rolnick et al and made public for upload to Kaggle [5].

Together, these datasets provide 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.

The energy dataset consists of 29 variables (columns, K) and 36064 measurements (rows, N), the excel data itself is 36065 rows as the first row contains the variable names. The weather features dataset consists of 17 variables (columns, K) and 178396 measurements (rows, N), the excel data itself is 178397 rows as the first row consists of the variable names. A summary of the variables and their units for each dataset is outlined by column in Appendix A. For both datasets the data is recorded hourly for four years, the weather features have five times the number of measurements of the energy dataset because the data for four years is obtained for five separate cities. The number of measurements for each city is equal to the of the energy dataset, and thus these datasets can be compared for each city individually.

The datasets needed to be preprocessed to be usable. First, the weather_features.csv was separated into five files, one for each city. To achieve this, a Perl script was written. Second, Excel's built-in features were used to remove empty columns and duplicate records. The processes were saved as VBA macro scripts. Lasty, MySQL was used to combine the weather features with data in the energy_dataset.csv file. An inner join was used to keep only data for which there was energy and weather records. The end result was five comma-separated value files. The scripts for performing this processing is available in a private GitHub repository [6]. An optional Perl script can separate datetime values into years, months, days, and hours. A summary of the variables and their units for the new merged datasets can also be found in Appendix A. All but the Madrid dataset was free of

major outliers. The Madrid dataset had an incorrect entry in two records of the actual price field. Electricity generation data was incorrectly placed in the price field, which This was discovered during ANN training and the two rows were manually removed.

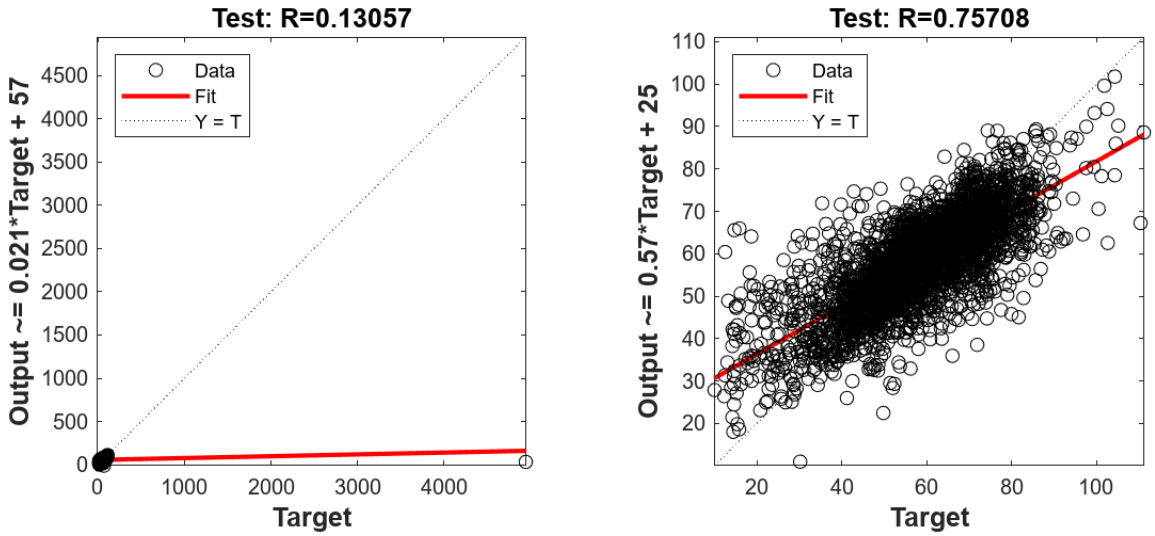


Figure 1: ANN test dataset regression for Madrid before and after removing outliers

Scatterplot matrices were created as a preliminary analysis of these combined datasets to discover if there were any baseline correlations between the weather variables and the total electrical demand and the price of electricity, as well as examine the distribution of theses variables. A matrix was made for each city, using the variables temperature, humidity, wind speed, wind degree, rain in one-hour, total cloud cover, total load, and actual price. The figures for each city can be found in Appendix B.

The matrices revealed that temperature, humidity, total load, and price generally all had normal distributions across the cities, although humidity in Bilbao had an exponential distribution. Wind speed tended to have a log-normal distribution, and cloud cover had what resembled a Bernoulli distribution. Rain in one-hour had little to no distribution, showing that rainfall must be very consistent.

The most obvious correlations displayed were a negative correlation between temperature and humidity in Valencia and Bilbao (though not the other cities), and a positive correlation between total load and price. The correlation between weather patterns and energy usage and price is less obvious. A very loose positive correlation between temperature and price can be seen in Barcelona only, and a very loose positive correlation between temperature and load in Bilbao, Madrid, and Valencia. While these correlations do not seem to be very strong, they serve as a starting point for further data analysis and provide insight into which variables to explore the connection between more in depth. Most notably, the correlation between temperature and various energy generation methods should be explored to reveal patterns between those variables, total load, and price of energy as well.

Further preprocessing methods such as PCA and PLS are utilized to reveal more correlations and predictors within this dataset.

Problem Statement

The goal of this data analysis is to predict energy load and energy price using datetime, weather, and/or energy generation data. PCA will be performed on the individual datasets and on the combined datasets to identify important variables and potential correlations. This will give some guidance when developing a PLS to predict energy load and energy prices.

Preliminary Results

Individual Datasets

PCA

The first step was performing PCA on each dataset individually to identify the important variables and select which variables should be included in subsequent modelling and analyses. For the weather dataset, PCA with 3 components was performed on all numerical variables. Score-loadings plots of the variables are shown in Figure 2. From this plot it was concluded that temperature, minimum temperature, and maximum temperature are strongly correlated and thus only temperature is required in future models. The variables close to the origin are year, day, city, 3hr rain, and 3hr snow. These variables have a very small impact on the scores. Variables that have a greater impact on the scores are wind angle and wind speed, temperature, humidity, cloud cover, month, and rain 1h. The R^2 of the 3 component PCA is 0.3334. R^2 does not improve when additional components are added. Most of the scores are tightly clustered about the origin, but there are some outliers that represent extreme weather days.

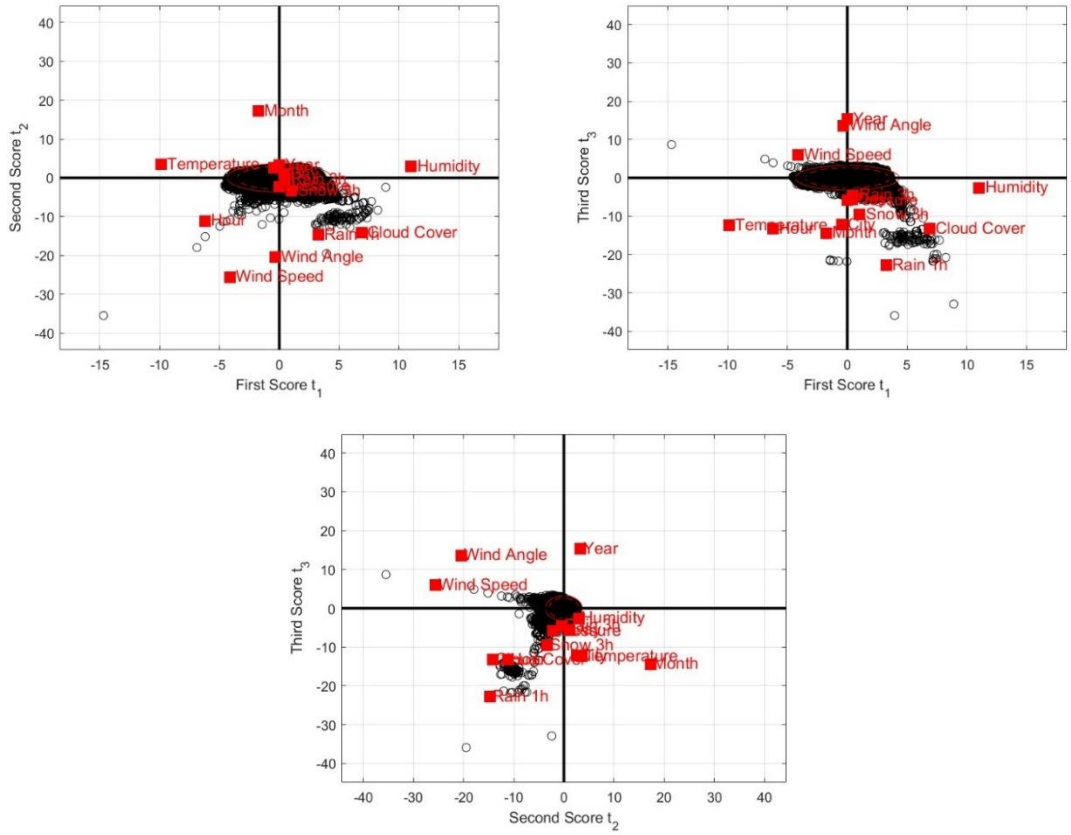


Figure 2: Score-loadings plots of the first three components for PCA on the weather dataset

PCA was also run on the energy dataset. A score loadings plot of the first two components is shown in Figure 3. No obvious conclusions can be drawn from the plot. Most variables contribute to the first component with Fossil-based generation having the largest loadings. The variables that have a greater effect in the second component are Biomass, Waste, Other, and Other Renewable energy generation. A plot of R^2 and Q^2 as a function of number of components is shown in Figure 4. The PCA model only becomes effective at explaining the data when more than 16 components are used. To attempt to obtain a more useful model, the PCA was run again on the energy dataset without datetime information. The score loadings plot and R^2, Q^2 plot are shown in Figure 5 and Figure 6. This model gives a clearer picture of the relationship between different energy generation as all fossil-based generation methods influence the first component. It can also be seen that beyond 16 components, additional components do not explain any additional variance in the model as Q^2 sharply decreases.

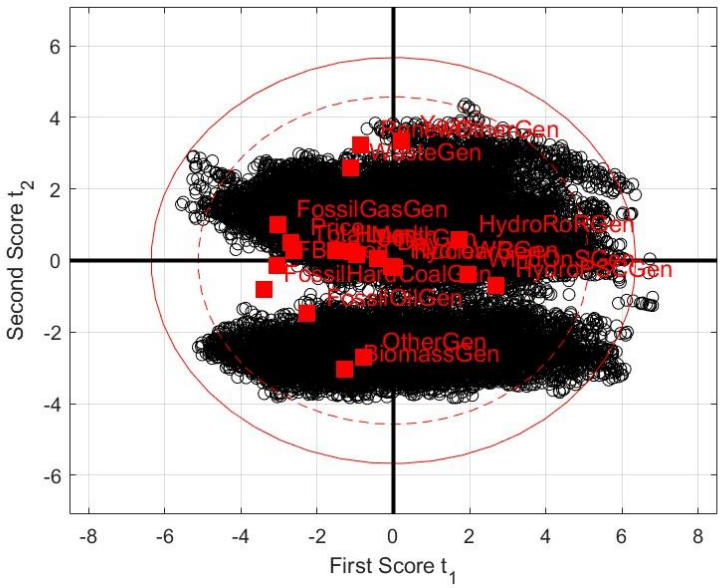


Figure 3: Score-loadings plot of the first and second component for a PCA run on the energy dataset with all variables

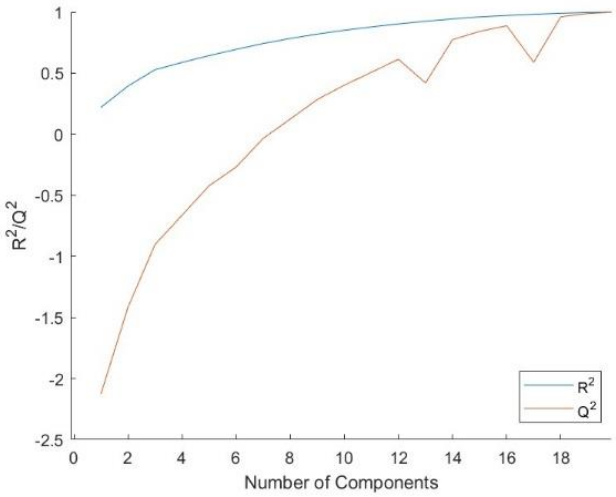


Figure 4: R^2 and Q^2 vs. number of components for PCA on the energy dataset with all variables

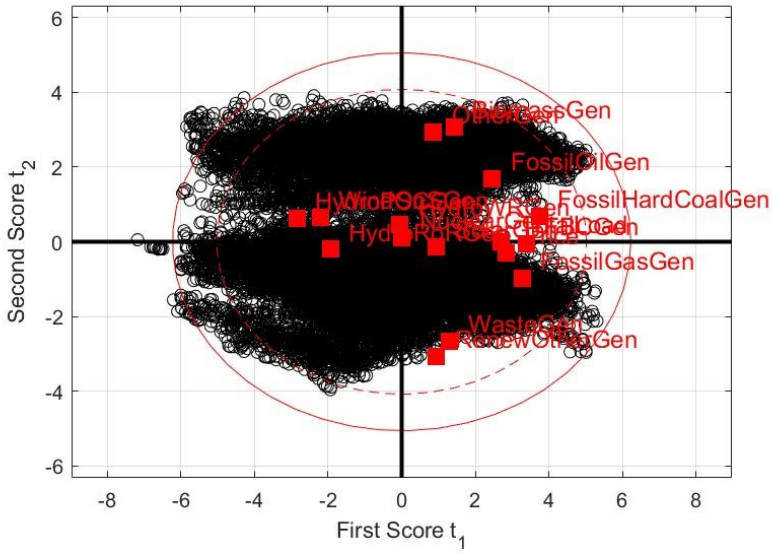


Figure 5: Score-loadings plot of the first and second component for a PCA run on the energy dataset without datetime variables

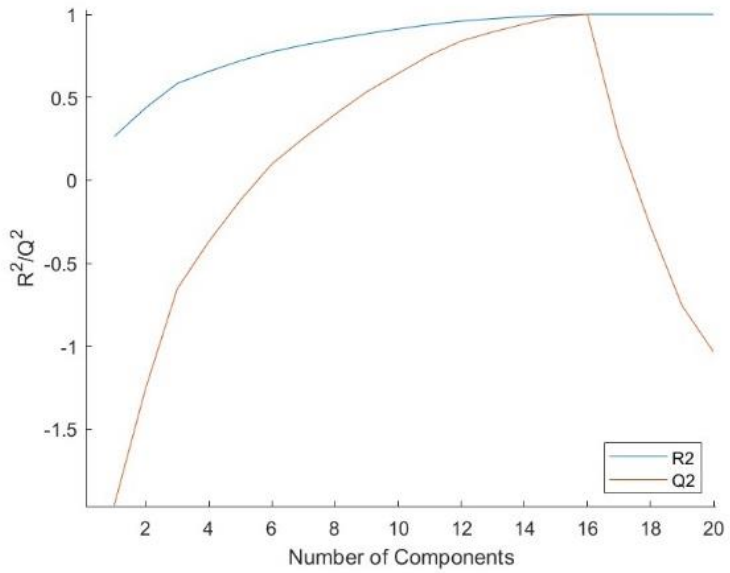


Figure 6: R^2 and Q^2 vs. number of components for PCA on the energy dataset without datetime variables

PLS

After determining which variables are important and looking for correlations between variables, a PLS was run on the energy dataset to predict the total energy load and energy price (Y) based on the datetime and generation variables (X). A PLS with 6 components was used and yielded an R^2 of 0.65. Beyond 6 components, R^2 showed marginal increases. Cross-validation was performed on the model with G=4 and all Q^2 values were approximately -0.35 which indicates very poor predictive power of the model. Figure 7 shows an SPE plot of the model and demonstrates many predicted y values with large error. A PLS was also run with datetime data omitted. This showed slight decreases in predictive performance.

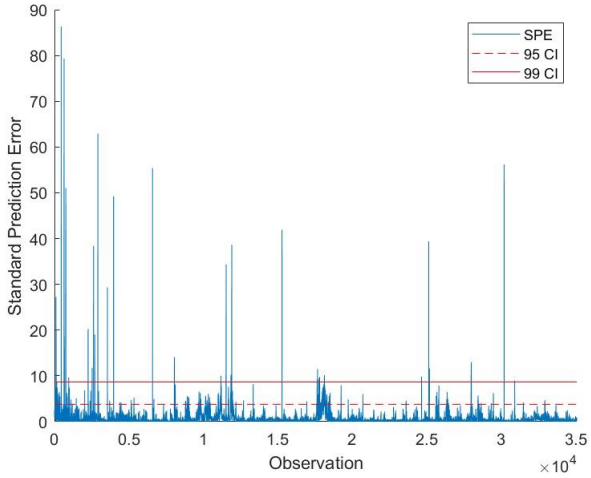


Figure 7: SPE plot for PLS on the energy dataset

ANN

ANN exploration on the data was performed on the individual sets. ANN models for all cities were similar. It was possible to fit a moderately good model predicting the price of energy based on the type of electrical generation. Windspeed is used to predict onshore and offshore wind power generation in. The regressions are presented in Figure 8. 20 hidden layers, 10% validation data, 10% testing data, and 80% training data was used to generate these models. ANN models for other input-output variables could not produce workable ANNs. Further processing including using past weather to predict future weather will be explored further.

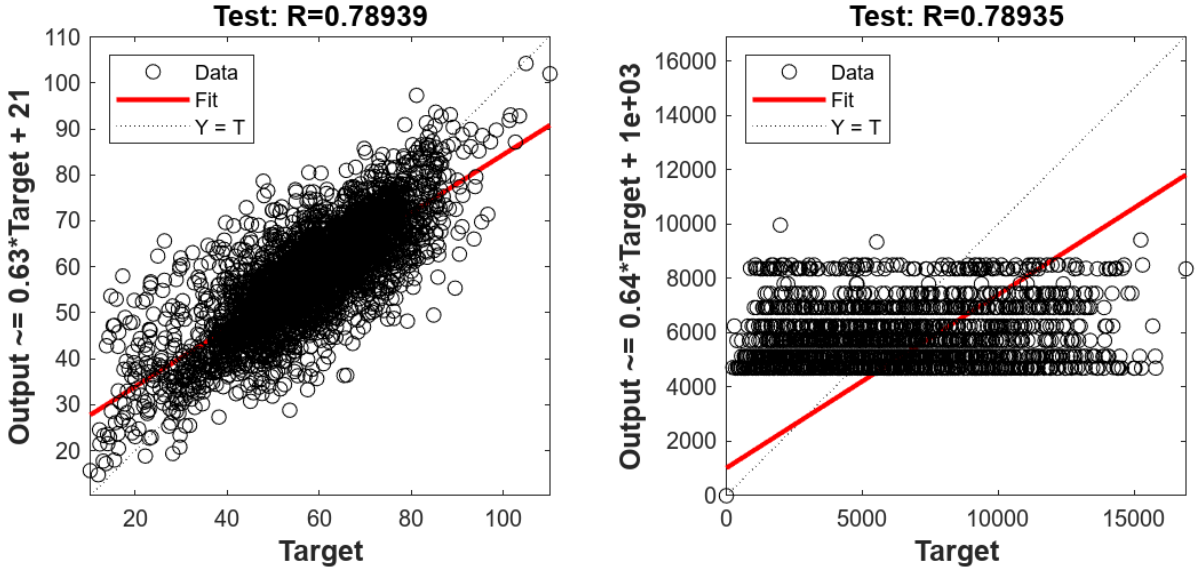


Figure 8: ANNs predicting actual price based on electricity generation source and wind power from windspeed

Combined Dataset

PCA

As a preliminary analysis of the correlation between weather and energy variables, a PCA was run on the combined weather and energy dataset for Barcelona. All subsequent analyses were performed only on Barcelona. Future work will be done on all cities. Figure 9 shows a score loadings

plot for a PCA run on weather and energy load and price data. Figure 10 shows the loadings plot of the same PCA for the first two components. In the first component, there is a negative correlation between humidity and energy load and price, however this is not a strong correlation because it becomes positive in the second component. In the second component, there is a negative correlation between wind angle and speed and energy load and price. Pressure and rain have minimal impacts on the scores and are not noticeably correlated with energy load and price. Outliers appear to be associated with high wind speed and low energy price. The R^2 of this PCA with 3 components is 0.48 and R^2 shows no improvement beyond 3 components. While this is not a good model, it does provide some insight into important variables and correlations.

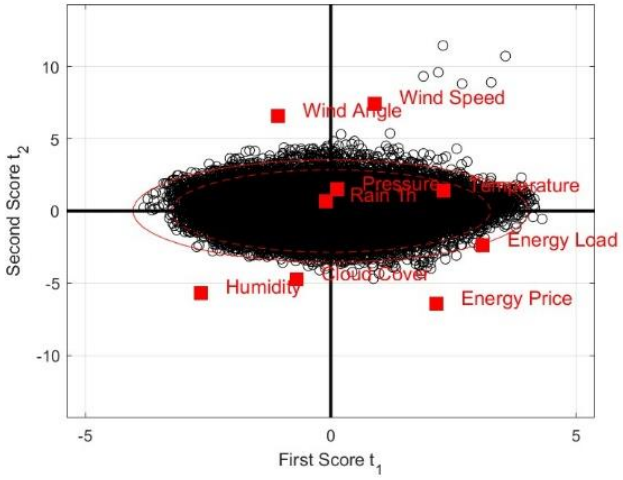


Figure 9: Score loadings plot of the first and second components for PCA on combined weather and energy data for Barcelona

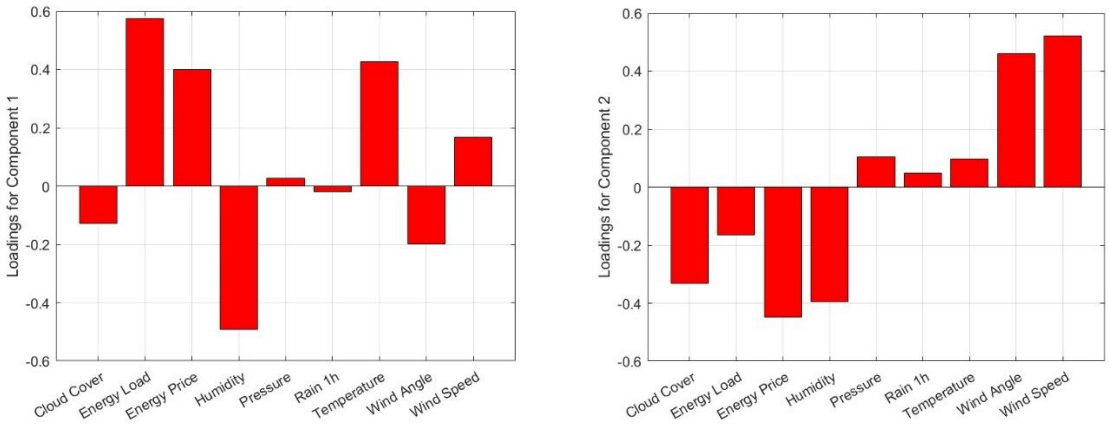


Figure 10: Loadings plot of the first and second components for PCA on combined weather and energy data for Barcelona

PLS

The primary goal of this data analysis is to predict energy load and energy price from datetime, weather, and/or energy generation data. First, PLS was done with X = datetime, weather, energy generation, and Y = energy load, energy price. Using 5 components gave an R^2 of 0.66. Any additional components show marginal increases in R^2. An SPE plot is shown in Figure 11. There is a significant outlier at observation 25083 corresponding to a data entry that is missing most energy

generation data, so this value was removed from the dataset and the PLS was rerun. This showed no significant improvements in R^2 or the other values of SPE as seen in Figure 12. There are many data points that lie outside of the 95% and 99% confidence intervals, Q^2 is negative regardless of how many components are added, and the R^2 is not impressive which indicates poor model predictive power.

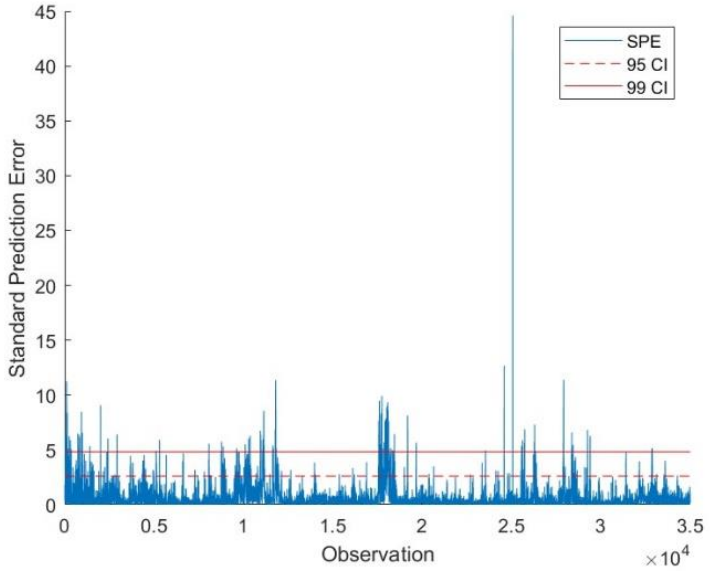


Figure 11: SPE plot for PLS on combined weather/energy dataset with X = datetime, weather, energy generation and Y = energy load, energy price

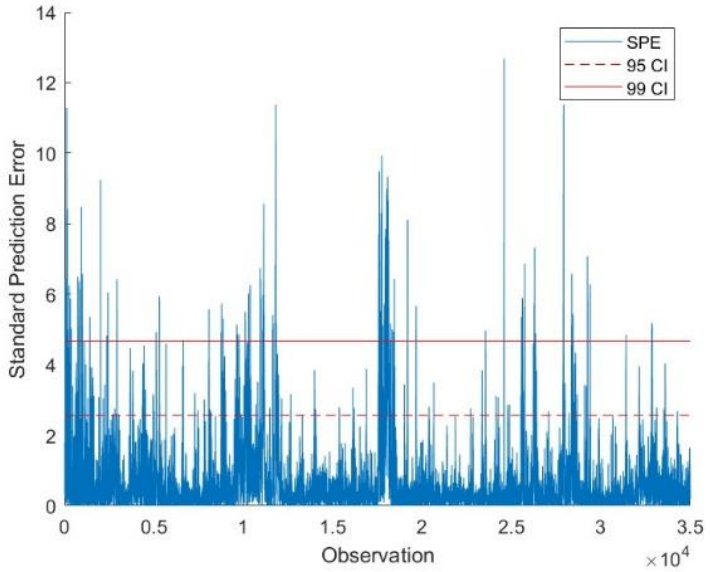


Figure 12: SPE plot for PLS on combined weather/energy dataset with X = datetime, weather, energy generation and Y = energy load, energy price with outlier removed

The influence of weather on energy load and energy price was also investigated by running PLS with X = weather and Y = energy load, energy price. This yielded a very poor model with an R^2 of 0.27 with 4 components and marginal improvements with additional components. An SPE plot is shown in

Figure 13 and indicates poor predictive power as many data points lie outside of the 95% confidence interval. Weather is therefore not sufficient to predict energy load and energy prices.

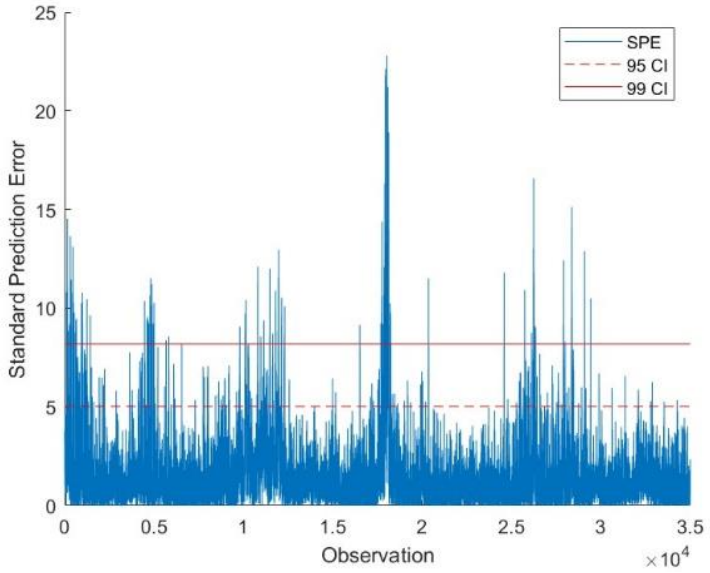


Figure 13: SPE plot for PLS on combined weather/energy dataset with X = weather and Y = energy load, energy price

Future Directions and Scope Limitations

Significant further work is required to develop a model capable of adequately predicting energy load and energy prices. Potential improvements may be made by building an ANN with weather and/or energy generation as the input variables. The number of variables in the energy dataset could likely be reduced to only important energy generation methods to reduce noise in the models. The effect of variable reduction will be investigated further. Once an effective modelling technique is established, it will be applied to all cities and, time allowing, the results can be compared between cities to identify any geographical differences.

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

[3] “ENTSO-e Transparency Platform,” ENTSO-e, https://transparency.entsoe.eu/dashboard/show (accessed Mar. 26, 2025).

[4] ESIOS Red Eléctrica, https://www.esios.ree.es/en/market-and-prices?date=18-03-2025 (accessed Mar. 26, 2025).

[5] Weather API - openweathermap, https://openweathermap.org/api (accessed Mar. 26, 2025).

[6] K. Snell, “4H03_group9.” March 18, 2025. Accessed: March 29, 2025. [Online]. Available: https://github.com/Kurt-Snell/4H03_group9

Appendix A – Variable and Unit Description of Datasets

Weather Features:

Columns 1-8

Variable Name:	Dt_iso	city_name	temp	temp_min	temp_max	pressure	humidity	Wind_speed
Explanation / Units:	Datetime index localized to CET	City name	Average temperature [K]	[K]	[K]	[hPa]	[%]	[m/s]

Columns 9-17

wind_deg	rain_1h	rain_3h	snow_3h	clouds_all	weather_id	weather_main	weather_description	weather_icon
Wind direction	Rain in last hour [mm]	Rain in last three hours [mm]	Snow in last three hours [mm]	Cloud cover [%]	Code used to describe weather	Short description of current weather	Long description of current weather	Weather icon code for website

Energy:

Columns 1-8

Variable Name:	time	Generation biomass	Generation fossil brown coal/lignite	Generation fossil coal-derived gas	Generation fossil gas	Generation fossil hard coal	Generation fossil oil	Generation fossil oil shale
Explanation / Units:	Datetime index localized to CET	[MW]	[MW]	[MW]	[MW]	[MW]	[MW]	[MW]

Columns 9-17

Generation fossil peat	Generation geothermal	Generation hydro pumped storage aggregated	Generation hydro pumped storage consumption	Generation hydro run-off-river and poundage	Generation hydro water resevoir	Generation marine	Generation nuclear	Generation other
[MW]	[MW]	[MW]	[MW]	[MW]	[MW]	[MW]	[MW]	[MW]

Columns 18-26

Generation other renewable	Generation solar	Generation waste	Generation wind offshore	Generation wind onshore	Forecast solar day ahead	Forecast wind offshore day ahead	Forecast wind onshore day ahead	Total load forecast
[MW]	[MW]	[MW]	[MW]	[W]				Forecasted electrical demand

Columns 27-29:

Total load actual	Price day ahead	Actual price
Actual electrical demand	Forecasted price [EUR/MWh]	[EUR/MWh]

Merged

Columns 1-7

Variable Name:	dttm	temp	temp_min	temp_max	pressure	humidity	Wind_speed
Explanation / Units:	Datetime index localized to CET	Average temperature [K]	[K]	[K]	[hPa]	[%]	[m/s]

Columns 8-16

wind_deg	rain_1h	rain_3h	snow_3h	clouds_all	weather_id	weather_main	weather_description	weather_icon
Wind direction	Rain in last hour [mm]	Rain in last three hours [mm]	Snow in last three hours [mm]	Cloud cover [%]	Code used to describe weather	Short description of current weather	Long description of current weather	Weather icon code for website

Columns 17-23

Generation biomass	Generation fossil brown coal/lignite	Generation fossil coal-derived gas	Generation fossil gas	Generation fossil hard coal	Generation fossil oil	Generation fossil oil shale
[MW]	[MW]	[MW]	[MW]	[MW]	[MW]	[MW]

Columns 24-31

Generation fossil peat	Generation geothermal	Generation hydro pumped storage consumption	Generation hydro run-off-river and poundage	Generation hydro water resevoir	Generation marine	Generation nuclear	Generation other
[MW]	[MW]	[MW]	[MW]	[MW]	[MW]	[MW]	[MW]

Columns 32-39

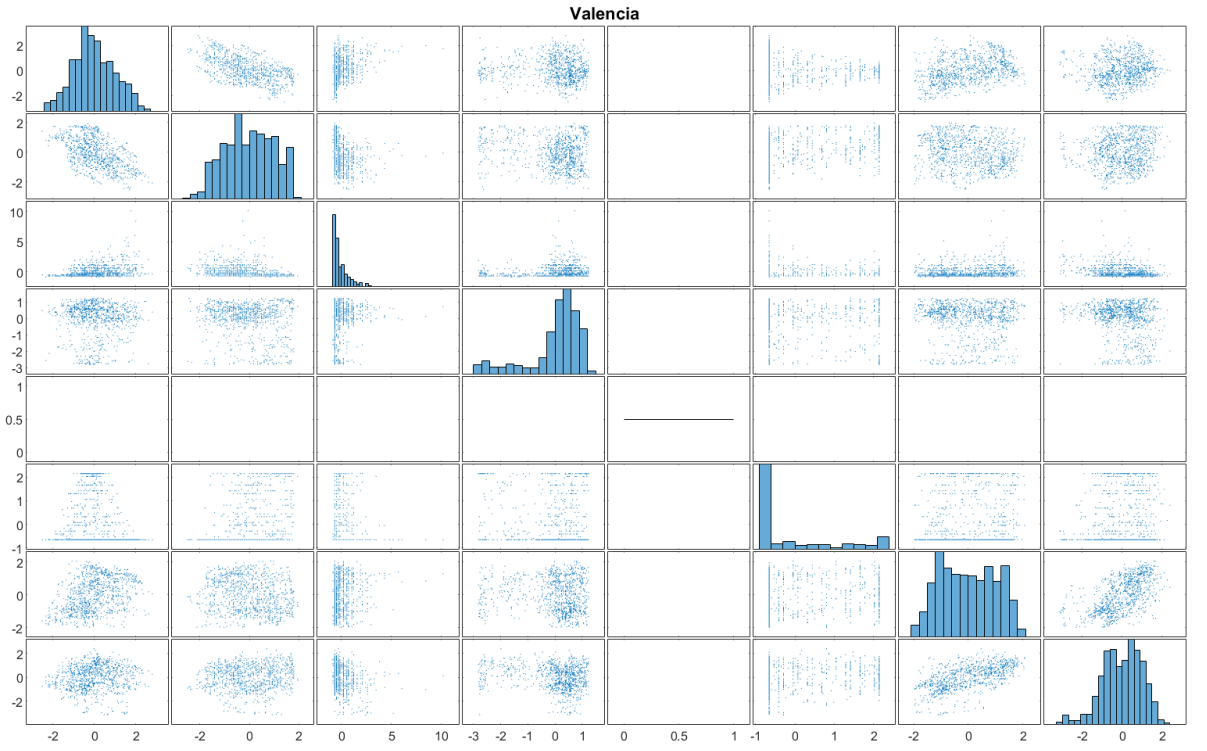
Generation other renewable	Generation solar	Generation waste	Generation wind offshore	Generation wind onshore	Forecast solar day ahead	Forecast wind onshore day ahead	Total load forecast
[MW]	[MW]	[MW]	[MW]	[W]			Forecasted electrical demand

Columns 40-42

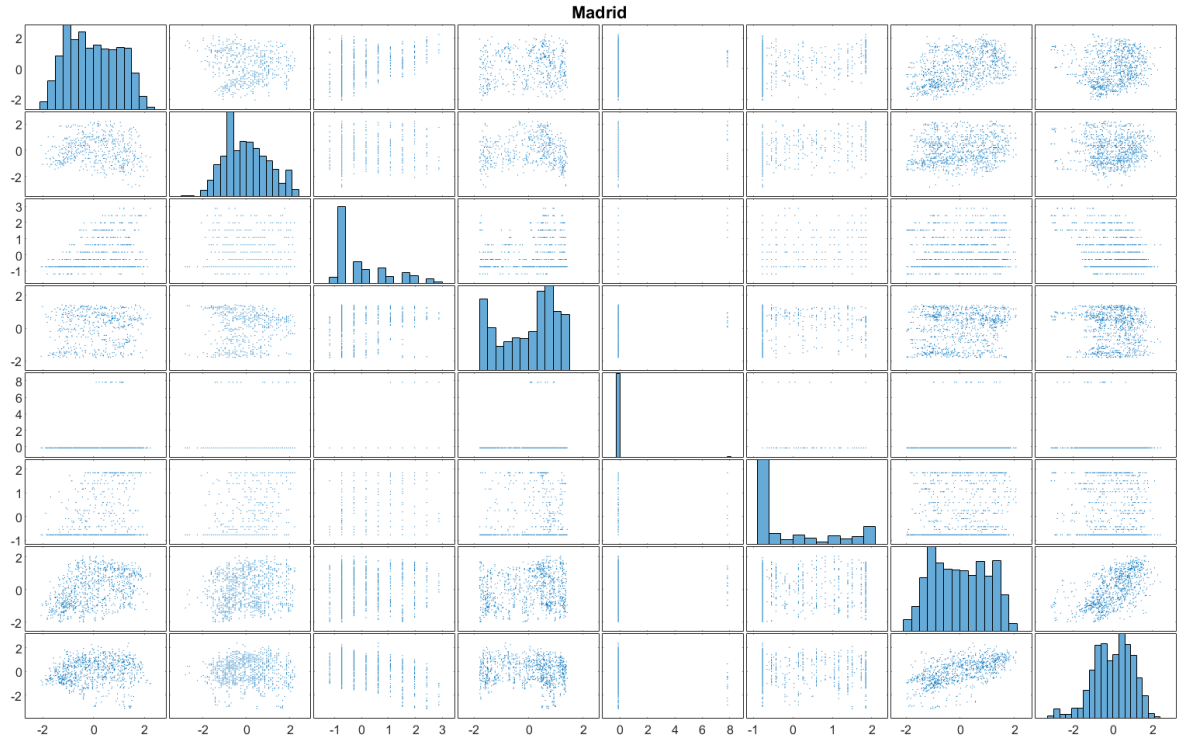
Total load actual	Price day ahead	Actual price
Actual electrical demand	Forecasted price [EUR/MWh]	[EUR/MWh]

B – Scatterplot Matrices per City

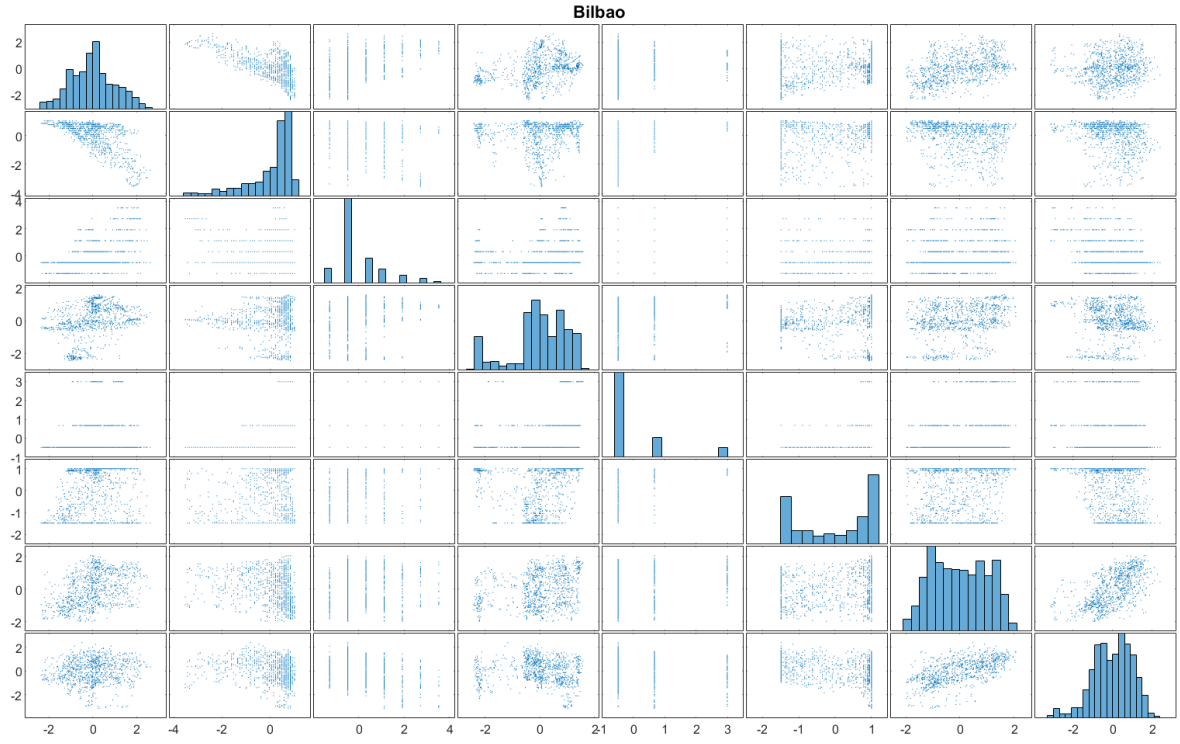
Valencia



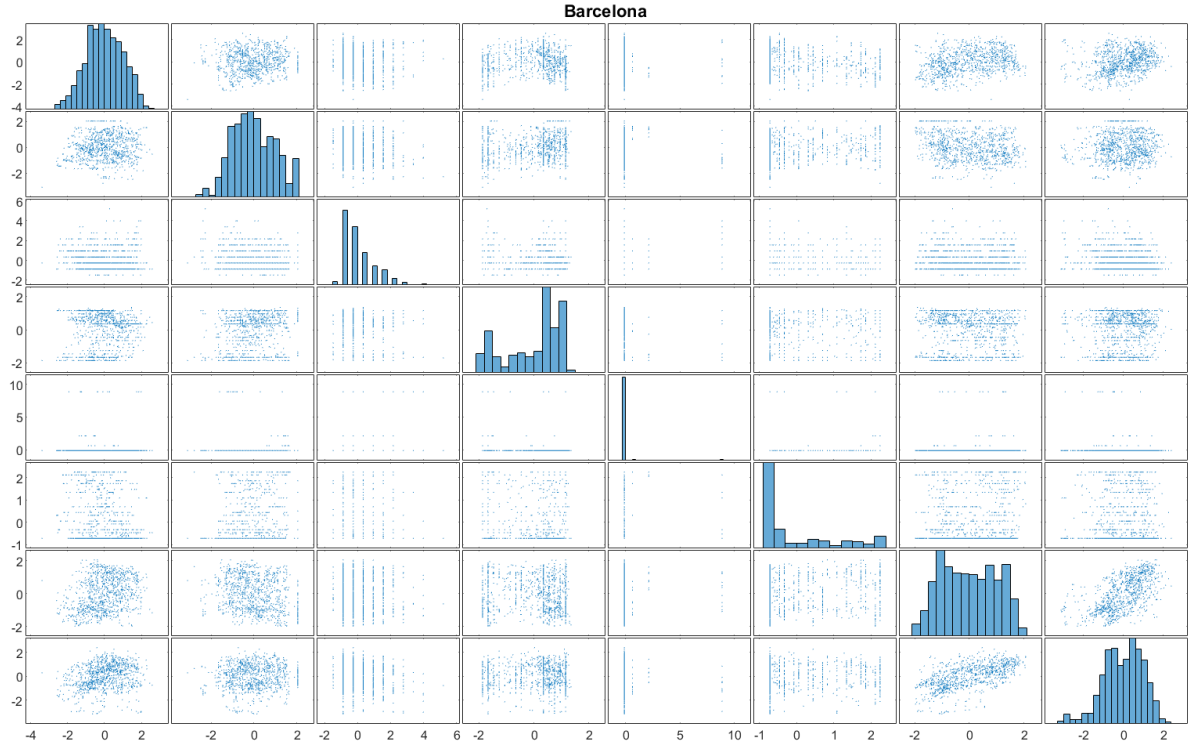
Madrid



Bilbao



Barcelona



Seville