- **Title**: *Multiple Linear Regression Model to Predict Electricity Demand of Five Cities in Spain*
- **Names**: *Keith Oxley, Shailesh Wasti, Tzu-En Chen, Yuxiao Zhu*

Abstract

The ability to estimate of electricity demand pattern allows electric grid operator design the market that serves the interest of people. This report models multi-linear and tree based regression using temporal and weather attributes. We performed the analysis for five cities in Spain over the four-year time period (2015-2018) all in hourly resolution. Assessment of five reduction methods yielded the p-value reduced model being the preferred from the MLR analysis for all the five cities. However, the RMSE and mean error (predicted vs actual) values were found to be double those of the tree regression model approach (91.2 MWH and 3.58% VS 51.7 MWH and 1.66%, respectively). Given the same condition, Tree model is preferred over MLR model. As a part of Tree analysis, we calculated the optimal number of branches in random forest model by looking at mean errors of accuracy threshold of 95%, which in turn significantly limited over fitting of the model. With reference to the Tree analysis importance plots, the three temporal variables coupled with daytime time band dummy variable have the greatest influence on energy demand prediction.

Introduction

The US power industry was deregulated in the late 1990s [1]. A non-profit entity called Independent System Operator (ISO) is entirely responsible for balancing the supply and demand in the area. Electricity market is unique in the sense that any imbalance in the power for time as short as 5 seconds has severe consequence. Any difference in supply and demand will result in the frequency excursions in the grid, and prolonged excursions lead to blackouts. The challenge, thus, is the prediction of load demand as accurately as possible.

Multiple Linear Regression (MLR) is a statistical technique that uses explanatory variables to predict the response variable(s) [2]. The model is subject to MLR if it holds four assumptions: (a) there should be a linear relationship between explanatory variables (independent variables) and response variable (dependent variable), (b) the residuals should be independent, (c) the residuals should have constant variance also known as homoscedasticity, and(d) finally, the residual of the model should be normally distributed [2].

It is in our best interest to introduce auto-correlation and partial auto-correlation measures. The auto-correlation measures the temporal correlation between the elements of a time series separated by a given number of time lags. In auto-correlation, $AC(p)$ denotes the Pearson correlation coefficient between time series data X_t and time shifted data X_{t-p} [3]. In partial auto-correlation, $PAC(p)$ measures the direct correlation between time series data X_t and time shifted data X_{t-p}. Partial is different in such it ignores the effects from previous lag periods.

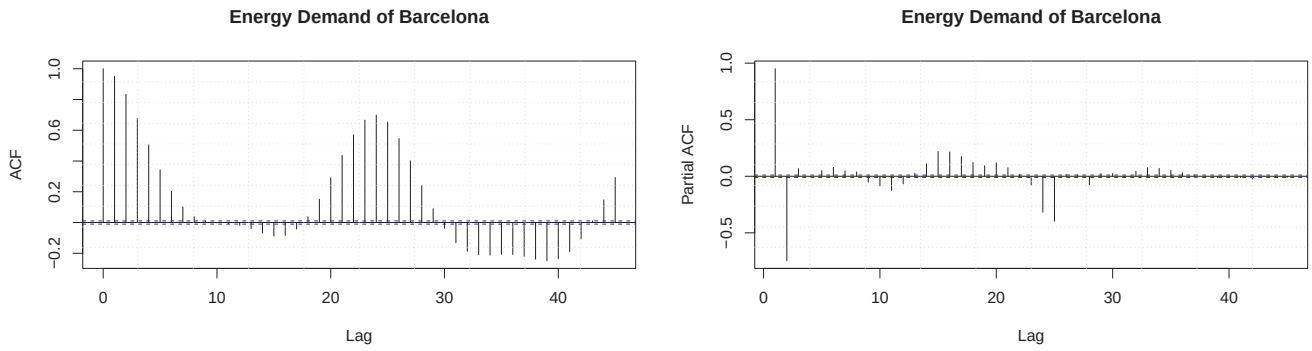

Figure 1: ACF and PACF Plot for Energy Demand of Barcelona

Fig. 1 shows the auto-correlation and partial auto-correlation plot for the city Barcelona. The plot shows that there is strong correlation between the past and present values of data sets. Observe that data point in position for partial correlation are $(E_1, E_2, E_{24}, E_{25}) = (0.95085, -0.7465278, -0.3161943, -0.3967825)$, where E is energy demand for given hour. This is an interesting observation as the demand of electricity depends on the demand last day in that hour and immediate past demand. For the purpose of the multi-linear regression analysis, the three largest (on an absolute value basis) temporal variables (E_1, E_2 and E_{25}) were brought into the model.

The contribution of this report can be summarized as below:

(i) Model weather attributes and temporal variables as explanatory variables to predict energy demand

(ii) Auto-regression and partial auto regression model to incorporate temporal variable in the regression model

(iii) Granularity of data is hourly, and we perform the analysis for 4 year and 5 cities in Spain

Data Source: The link to the data source is referenced in footnote. We use the data of 4 years, all in hourly resolution. This is 24 hours/(day-city) $\times$ 365 days/year $\times$ 4 years $\times$ 5 cities + 24 (2016 is a leap year) = 175224 Rows. The **Quantitative Variables** are (6): Ambient temperature,Humidity, Pressure, Wind Speed, Rain Duration, Snow Duration, and prior period actual energy demand obtained from partial auto-regression modeling and the **Categorical Variables** are (4): Day/Night, (billing) Time band (Peak/Mid-Peak/Off-peak, Season, and (12 descriptive) Weather designations. Note that categorical variables have sub groups.

Data Cleaning and Feature Engineering

The data from Kaggle had missing data, and duplicate data. We used R to clean those data points: we removed any duplicate data (keeping just the first of all the duplicate rows). We checked historical weather attributes (reference link in appendices), and removes any data that would was erroneously present in the data set. We selected the columns of our interest, and made indexed categorical variables. For example: "clear", "clouds", "drizzle", "dust","fog","haze","mist","rain","smoke","snow","squall","thunderstorm" are numbered from 1 to 12 in the same order so as to quantify weather.

Fig. 22 in Appendix shows the range of weather attributed to kept in our data frame for Barcelona: rain duration: $0 - 12.5$ millimeters, humidity: $0 - 100\%$, wind speed $0 - 15$ m/s, and pressure $0 - 1500$ mbar.

The hourly energy demand data from Kaggle was that for the overall country of Spain (not by autonomous region or principal city). This was proportioned to the five cities of investigation (Barcelona, Bilbao, Madrid, Seville and Valencia) based on population and on industrial level using a 60/40 split. Example of corresponding weight factor derivation for Barcelona (that for the other cities plus base data references can be found Table 4 in the appendix):

$WeightFactor_{Barcelona} = 0.6 \times (\frac{0.35}{46.67}) + 0.4 \times (\frac{10.4\%}{100\%}) = 0.0461 = 4.61\%$

The entry model for going into both the MLR and Tree-based regression modeling efforts (coefficient decriptions can be found in Table 10 in the appendix):

$$EnergyDemand = \beta_0 + \beta_1(E_1) + \beta_2(E_2) + \beta_3(E_{25}) + \beta_4(Temp) + \beta_5(Humidity) + \beta_6(Pressure) + \beta_7(WindSpeed) + \beta_8(RainDuration) + \beta_9(SnowDuration) + \beta_{10}(DayNight) + \beta_{11}(TimeBand) + \beta_{12}(Season) + \beta_{13}(WeatherMain)$$

MLR Analysis

Before we start to execute the Multi-Linear Regression (MLR) analysis, we check the validity of assumptions including linearity, heteroscedasticity, normality leverage points, influential points, and multicollinearity issue. In the beginning, we discover that a serious multicollinearity issue exists in the data set, thus, the E_2 variable was removed to address the issue.

While executing the model selection, the P-value selection method leads us to drop predictor, rain duration, then we build a reduced model to compare with the full model. According to the Anova output we do not reject the null hypothesis, which means we favor over reduce model Furthermore, we use the same step to compare each selection method including P-value, Best Fits, AIC forward, Cross Validation and Ols step with the full model. Finally, P-value selected model passes the Anova test and has the highest Adjusted R-square.

The final MLR regression model for Barcelona is:

$$EnergyDemand_{Barcelona} = -2804.181 + 0.885(E_1) + 0.121(E_{25}) + 7.162(Temp) - 0.638(Humidity) + 0.734(Pressure) - 1.963(WindSpeed) + 65.425(DayNight_2) - 17.155(TimeBand_2) + 62.623(TimeBand_3) - 72.999(Season_2) - 29.911(Season_3) + 17.237(Season_4) - 2.312(WeatherMain_2) - 13.149(WeatherMain_3) + 156.902(WeatherMain_4) + 16.93(WeatherMain_5) - 43.54(WeatherMain_7) + 15.811(WeatherMain_8) + 118.504(WeatherMain_{10}) + 3.57(WeatherMain_{12})$$

Based on the Correlation spectrum 23 in Appendix , observe that E_1 temporal regression predictor has the highest correlation with energy-demand while E_2 temporal regression predictor has a 0.83 correlation with the energy-demand. However, due to a serious multicollinearity issue, we dropped it from the model. Furthermore, the rain-

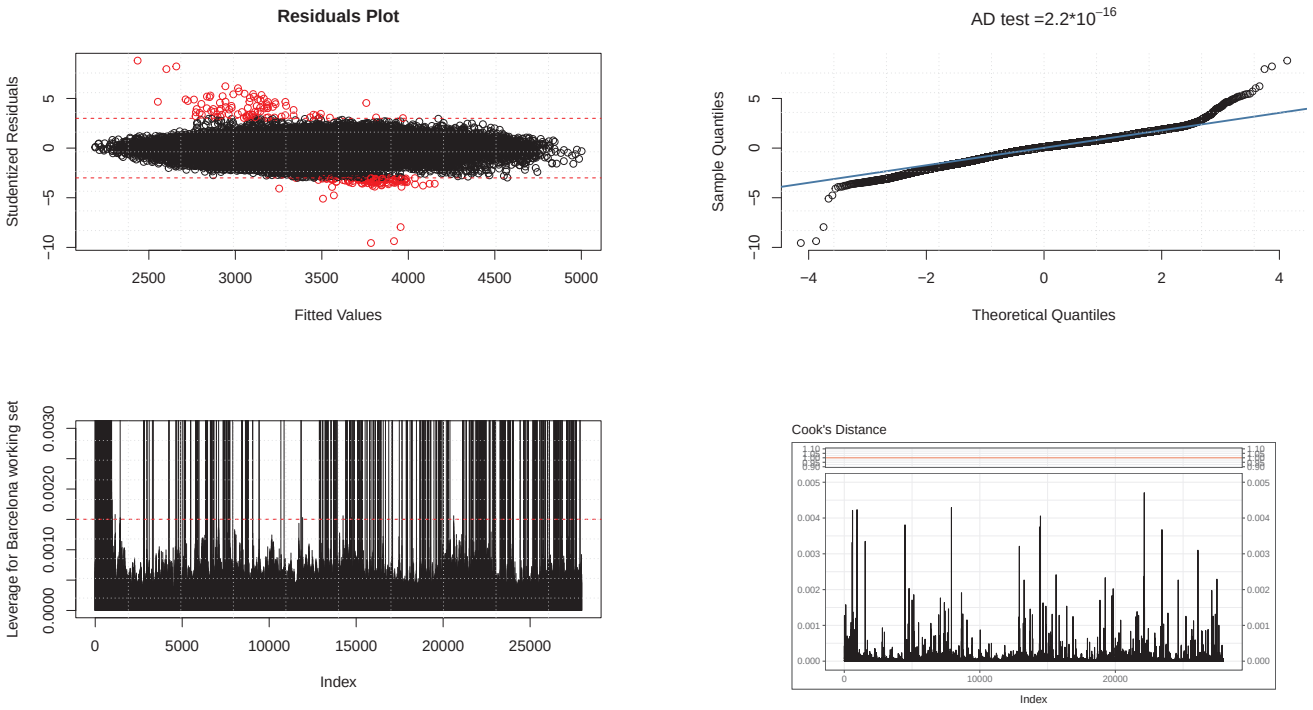

Figure 2: Model Assumption Check for Barcelona

duration predictor has the least correlation with the energy-demand. Therefore, we drop it according to the p-value output.

According to the residual plot 2, the model doesn't violate the linearity, heteroscedasticity doesn't exist and there is only 4.09 % of outliers in the data set thus not material. Moreover, based on the Norm Q-Q Plot, the normality assumption is not violated. Finally, there are no influential points in Cook Distance Plot.

1	2	3	4	5	Final
				OLS-step	P-Value
			Cross-valid	P-Value	
		AIC	P-Value		
	Bestfits	P-Value			
P-Value	P-Value				
Full					

Reduction approach	Variables	Coefficients	Adjusted R^2
Full	12	22	0.915555
P-Value	11	21	0.915552
Bestfits	7	9	0.915164
AIC	12	22	0.915555
Cross-validation	10	20	0.915511
OLS-step	11	21	0.915552

Table 1: ANOVA Model Comparison Illustration (left) and Illustrative Results from Barcelona MLR Analysis (right)

Tree-Based Analysis

Besides MLR, we explored whether there are tree-based models that are more appropriate for our data. Random Forest is a tree-based machine learning algorithm that uses the power of multiple decision trees to make decisions. Compared to decision trees, random forests not only reduce the impact of outliers, but also reduce the possibility of overfitting, which is a better choice for our data sets. We have created models each utilizing two different packages base on RF: randomForest and ranger. We attempted to compare these two methods to choose a more suitable model.

Before building models, in order to facilitate our subsequent modeling, we convert categorical variables (Day, Period, Intensity, Location, Weather) to type category in that they were initially numeric in the data. For randomForest, we first find how many trees in the forest that give us the smallest mean square error (MSE) after a seed set for duplicating the results in the future. The package randomForest is used to run our first model with the working data. And then we checked the MSE and system time of the whole forest for the comparison later. Meanwhile, we did the same building and checking for the model with package ranger. At the same time, we want to find the most

influential predictors in our model. Take Barcelona for instance, function varImpPlot to show the top 5 significant variables (the usage of 1 hour ago, day or night, the usage of 2 hours ago, the usage of 25 hours ago and season). We can also conduct more in-depth studies in the future on the effectiveness of these variables

We compared the differences for the five cities to select models with less MSE. According to the comparison Table 2, We find that the MSE of randomForest model is more than ten percent less than that of ranger, which is over 5 %, we can consider it as a statistically significant difference. While ranger is a simplified version of randomForest, it ran quicker and there were different levels of speed increase depending on the device we used. In this project, we focus more on the accuracy, but in practice we may sacrifice a small part of the accuracy for the sake of higher efficiency.

City	Full Random Forest (RF)	Tree Count	Ranger (Simplified RF)	Difference
Barcelona	88.44	83	99.66	12.7%
Bilbao	35.12	70	39.54	12.6%
Madrid	61.33	78	67.15	9.5%
Seville	31.17	87	34.87	11.9%
Valencia	42.42	81	46.90	10.6%

Table 2: Root Mean Squared Error Comparison of Two RF Approaches (Full vs Simplified)

The prediction versus actual data of Barcelona is shown on the right of Figure 4. It can be found that our predictions are generally consistent with the facts. So we assume the RF model with randomForest package becomes a very competitive candidate for our final selection.

Comparison of Models

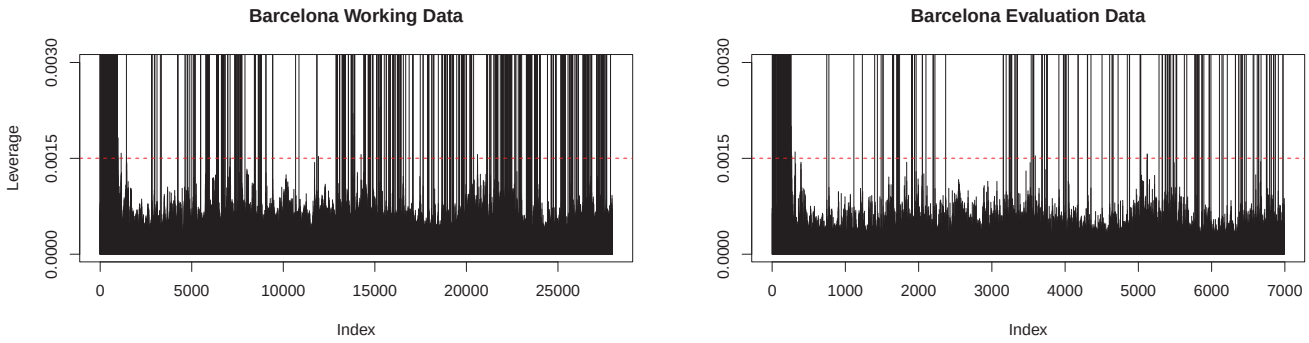

Figure 3: Barcelona Leverage Value Comparison

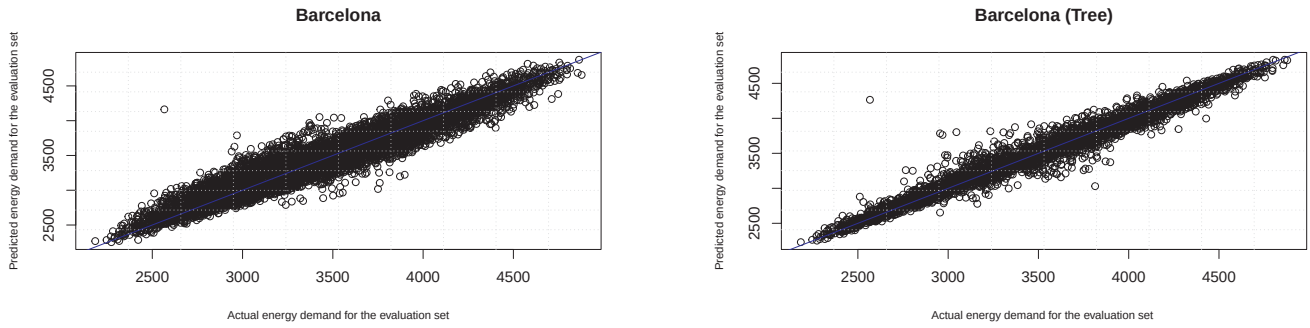

Figure 4: Predicted and Actual Energy Demand Comparison (MLR left, Tree right)

Following the separate but parallel MLR and Tree-based regression model development on the working data sets of the five cities in Spain (Barcelona first, then repeated for the other four; each started from a common, full regression model consisting of up to 14 combined numeric and categorical variables), each model was trained and then applied to to the evaluation data set parameters to predict the respective city's hourly energy demand. Performance assessments were then performed against the evaluation data set actual hourly demand for each city. During this process, comparison plots of the MLR (hat matrix analysis) evaluation data set leverage values were checked against those of the matching working data set to ensure alignment (as indicated in Table 3, all were found to be comparable); Fig. 3 shows this for Barcelona (that for the other four cities is included in the appendices). The overall performance of each modeling approach was then compared based on the resulting root mean squared error (RSME) for each city to determine which method was the best suited (lowest RMSE is best, thus the Tree approach is the preferred analysis method for all cities). Similarly, plots of the predicted vs actual energy demand were generated usig the resulting MLR and Tree models for each city's evaluation data set; Fig. 4 shows this for Barcelona that for the other four cities is included in the appendix). Table 3 summaries the overall findings of the performance comparison:

	MLR Leverage		MLR		Tree		RSME		
City	Working	Evaluation	Mean Error	Variables	Mean Error	Variables	MLR	Tree	Δ (vs Tree)
Barcelona	4.09%	4.26%	3.58%	11	1.63%	13	158.46	88.44	-70.02
Bilbao	6.66%	7.13%	3.64%	12	1.69%	13	61.98	35.12	-26.86
Madrid	6.76%	6.31%	3.57%	12	1.68%	13	107.04	61.33	-45.71
Seville	7.62%	7.60%	3.49%	11	1.65%	14	54.39	31.17	-23.22
Valencia	3.38%	3.35%	3.60%	12	1.67%	13	73.90	42.42	-31.48

Table 3: Performance Summary: MLR Leverage, Mean Error (Predicted vs Actual) and RSME Comparisons

As is indicated in Table 3 and illustrated in Fig. 4, the best regression modeling approach for Barcelona is the Tree-based. The same is true for the other four cities (Bilbao, Madrid, Seville and Valencia; comparable figures for these cities can be found in the appendices).

Discussion and Limitations

Discussion: (i) The P-value model was the preferred MLR output construct. (ii) There is very good performance of both the MLR and Tree based regression analysis models (mean errors are 3.58% and 1.66% for the MLR and Tree approaches, respectively, while their corresponding RMSEs are 91.2 and 51.7 MWh, respectively). (iii) In all cases, the Tree approach out performed the MLR modeling and thus is the selected approach. (iv) With reference to the Tree analysis importance plots (Fig. 17 ,18 ,19,20,21) in the appendix, the three temporal variables coupled with daytime time band dummy variable have the greatest influence on energy demand prediction.

Limitations: (1) In spite of best efforts, we were unable to obtain corresponding hourly energy demand by city over the four year period. Consequently, the overall country energy demand was allocated to city by a population and industrialization weighting factor. Given the geographic breadth of the five cities, weather in one city differs not only day to day to the other cities, but also hour by hour. Consequently, fine resolution of the actual association between a city's demand and given the local weather was lost; the country's hourly demand profile was simply scaled to each of the five cities (plus the remainder of the country) using the aforementioned 60/40 weighting factor. While the resulting analysis showed low predicted vs actual mean error percentages and root mean squared errors, appreciable improvement would be expected if the energy demand and weather data directly tracked. (2) Given the break between demand and actual weather, there was little improvement seen in adjusting population by year (the population percentage was fixed at 2015 levels for all four years). (3) While we were able to obtain industrialization levels for the autonomous region each city was in, it was assumed that these were co-located with each city. The land area of the Basque and Madrid autonomous regions (containing Bilbao and Madrid, respectively) are small relatively speaking, so it is not viewed that attributing the energy consumption by the industrial segments in those areas being co-located with each city is improper. However, the autonomous regions where the other three cities are located (Andalusia where Seville is, Catalonia where Barcelona is and Valencia where Valencia is) are all much larger (reference Table 5 in the appendix), thus viewing the industrial segments for those being co-located with the respective regional city probably introduces a material source of error.

References

[1] S. Arango, I. Dyner, and E. R. Larsen, "Lessons from deregulation: Understanding electricity markets in south america," *Utilities policy*, vol. 14, no. 3, pp. 196–207, 2006.

[2] S. Sheather, *A modern approach to regression with R*. Springer Science & Business Media, 2009.

[3] S. Makridakis and M. Hibon, "Arma models and the box–jenkins methodology," *Journal of forecasting*, vol. 16, no. 3, pp. 147–163, 1997.

Appendix

Energy Demand Data Preparation

Country level energy demand allocation - The following links were accessed to compile the population and industrialization columns in Table 4:

- Spain: The largest cities in 2015
- Production specialisation by autonomous region

City	2015 Population (million)	Regional Industrialization Percentage	60/40 weight factor
Bilboa	0.35	10.40	4.61%
Barcelona	1.60	25.10	12.10%
Madrid	3.14	9.80	7.96%
Seville	0.69	8.30	4.21%
Valencia	0.79	11.40	5.58%
Remainder	40.10	35.00	65.55%
Spain(Total)	46.67	100.00	

Table 4: City Allocation Basis for Overall Energy Demand

$WeightFactor_{City} = 0.6 \times (\frac{\text{City population}}{\text{Country population}}) + 0.4 \times (\frac{\text{City Industrialization Percentage}}{100\%})$
and $E_{City} = WeightFactor_{City} \times E_{Country}$ for any given hour over 2015-2018 timeframe.

Autonomous Region	City	Area (sq km)
Catalonia	Barcelona	32,114
Basque	Bilboa	7,234
Madrid	Madrid	8,028
Andalusia	Seville	87,268
Valencia	Valencia	23,255

Table 5: Autonomous Region Size Comparisons per Autonomous communities of Spain

Basis for cleaning Kaggle Weather Data Set by city:

- Average-Weather-in-Barcelona-Spain-Year-Round
- Average-Weather-in-Bilbao-Spain-Year-Round
- Average-Weather-in-Madrid-Spain-Year-Round
- Average-Weather-in-Sevilla-Spain-Year-Round
- Average-Weather-in-Valencia-Spain-Year-Round

Model Assumption Check for Rest of the Cities

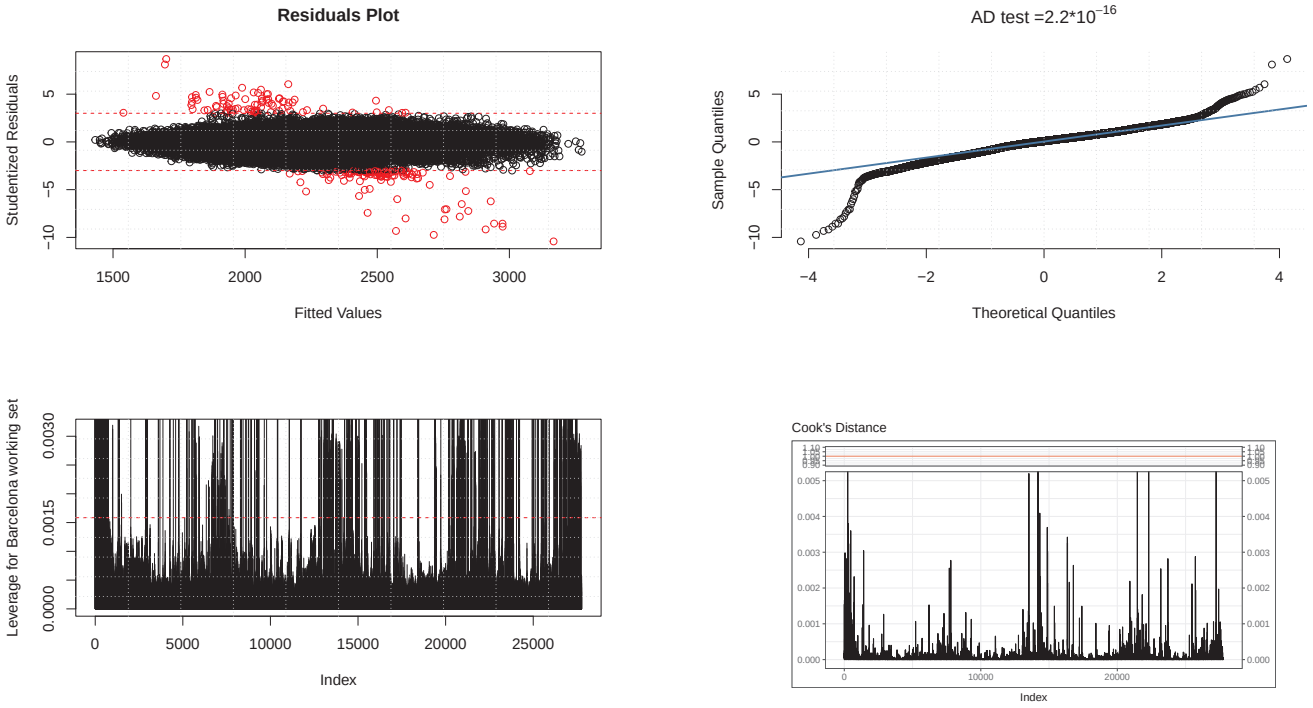

Figure 5: Model Assumption Check for Madrid

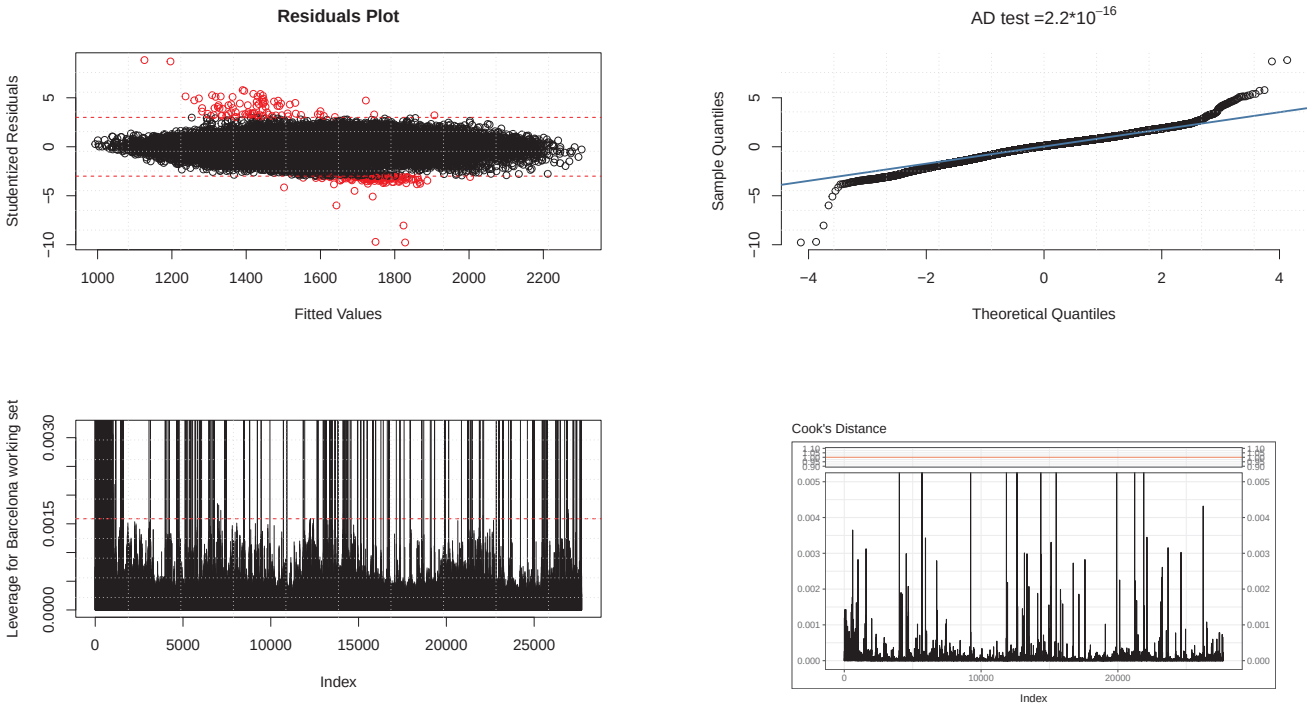

Figure 6: Model Assumption Check for Valencia

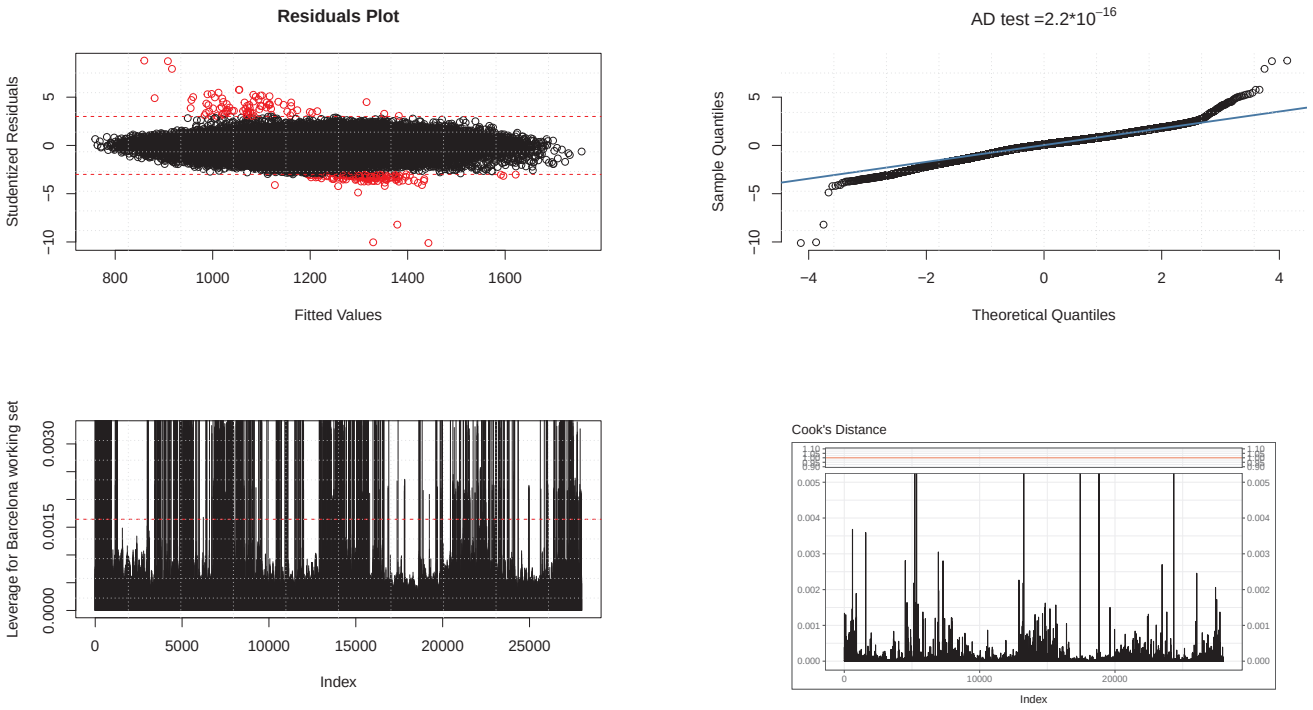

Figure 7: Model Assumption Check for Serville

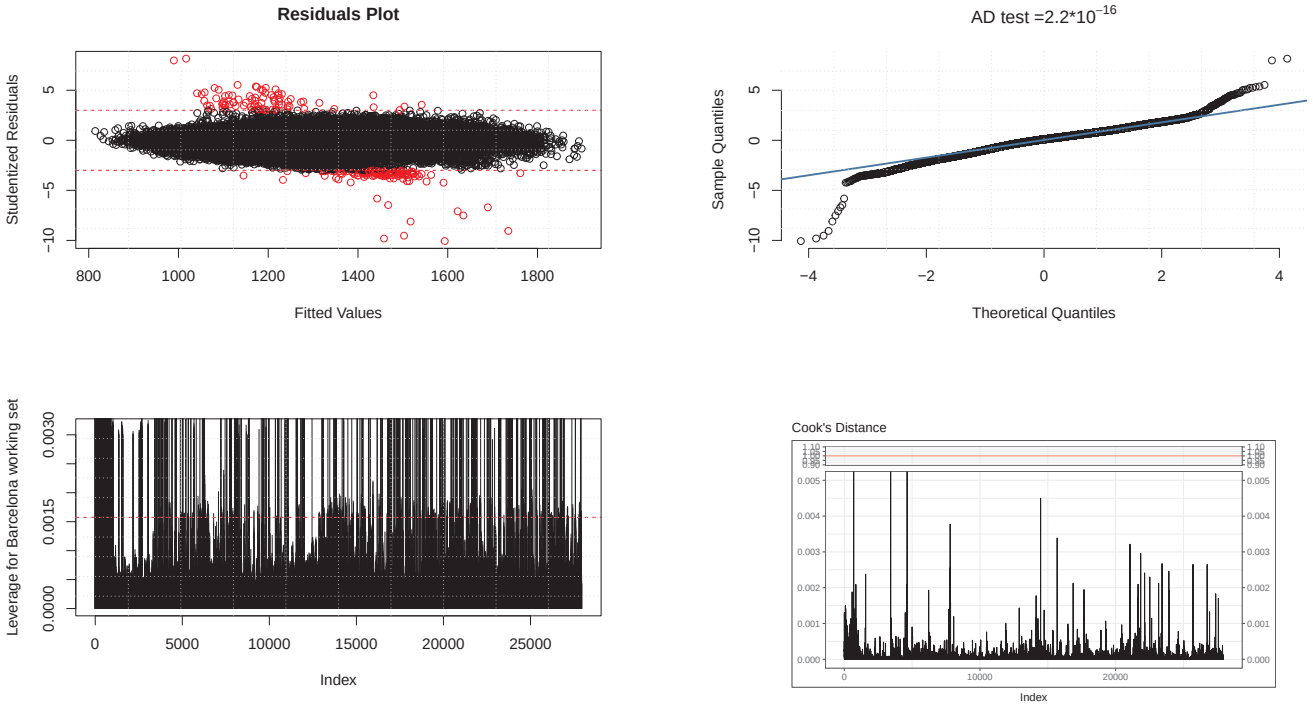

Figure 8: Model Assumption Check for Bilbao

Leverage Comparison for Rest of the Cities

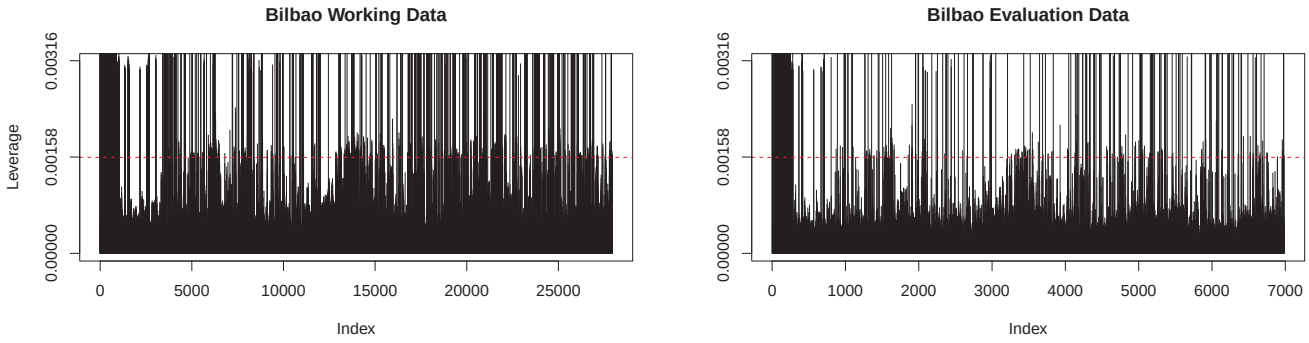

Figure 9: Bilbao Leverage Value Comparison

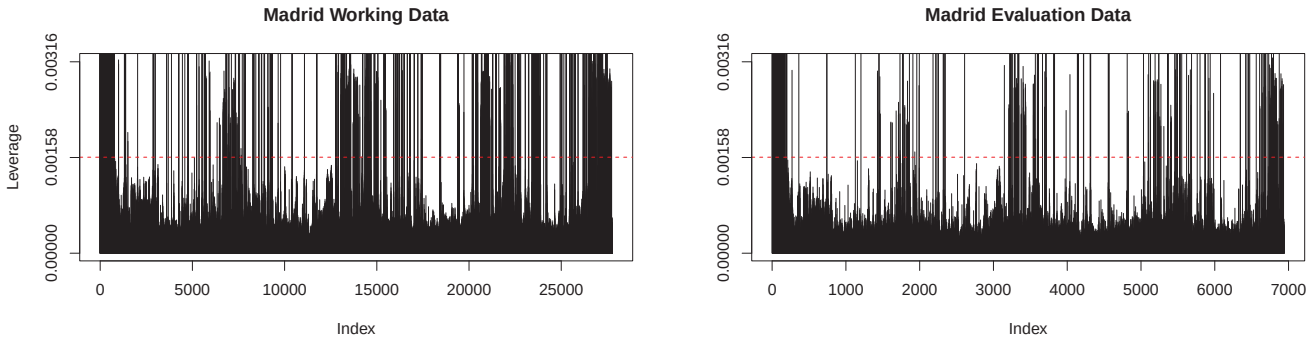

Figure 10: Madrid Leverage Value Comparison

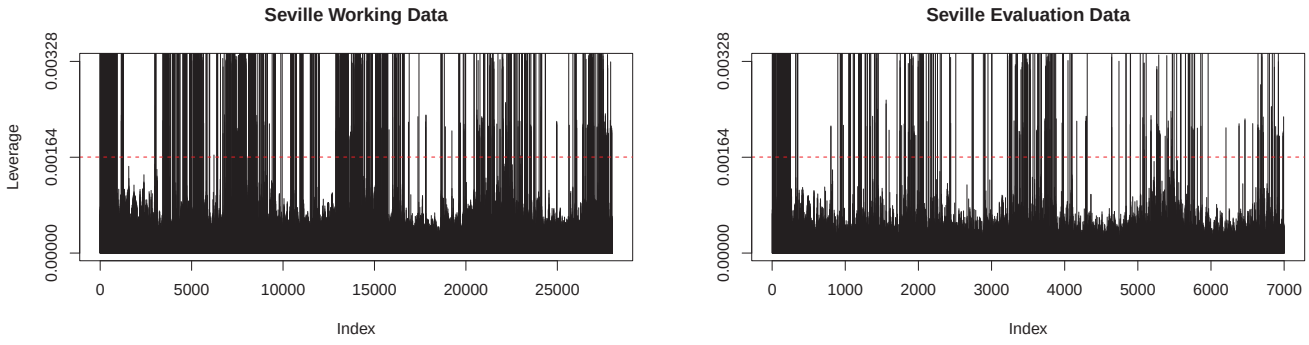

Figure 11: Seville Leverage Value Comparison

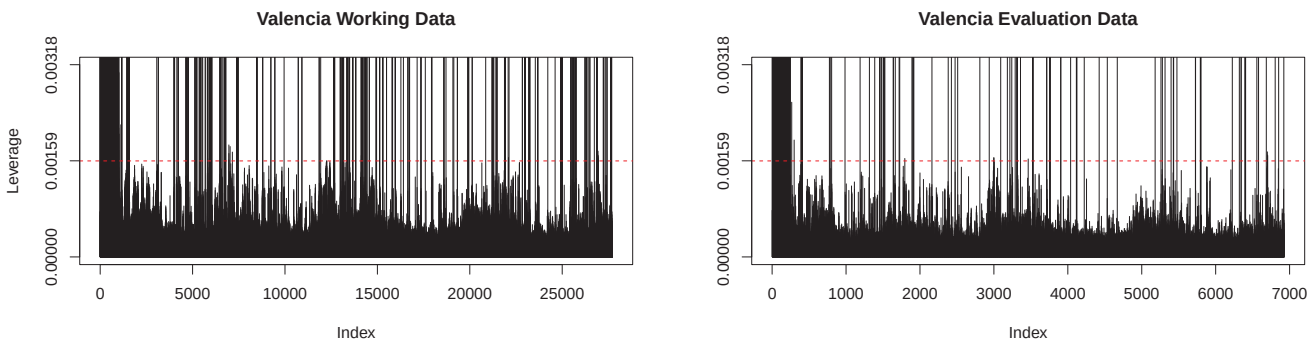

Figure 12: Valencia Leverage Value Comparison

Actual VS Predicted Energy Demand for Rest of the Cities

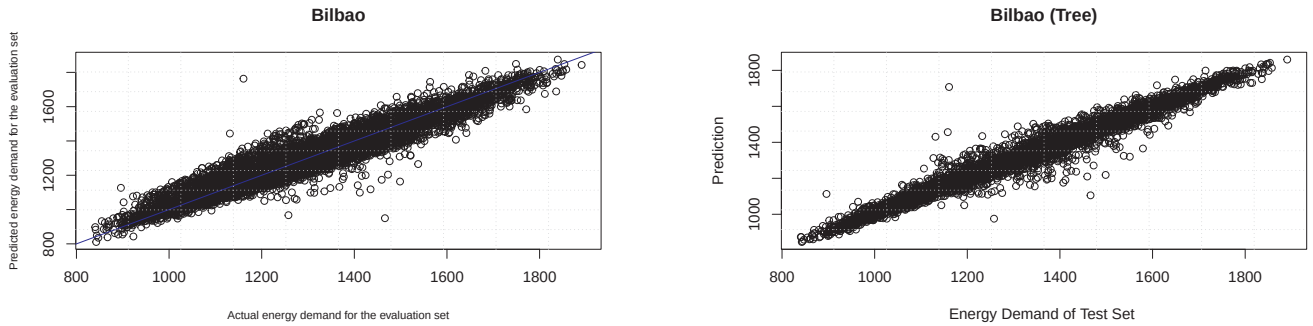

Figure 13: Predicted and Actual Energy Demand Comparison (MLR left, Tree right)

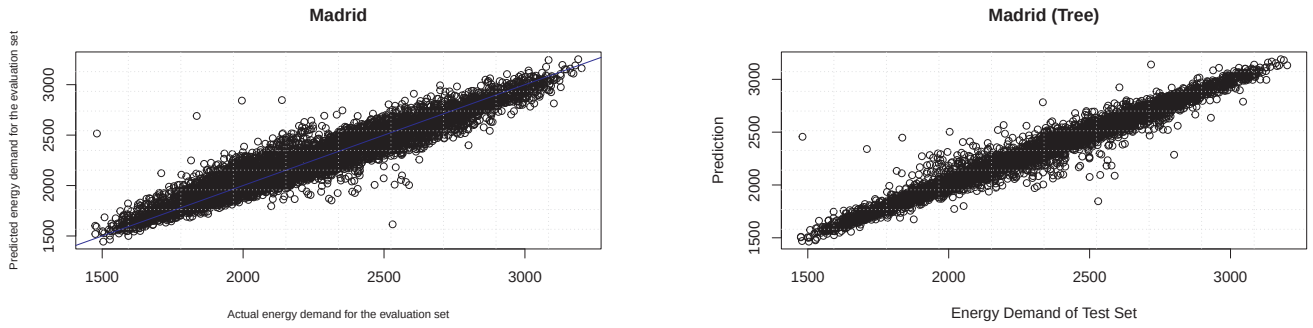

Figure 14: Predicted and Actual Energy Demand Comparison (MLR left, Tree right)

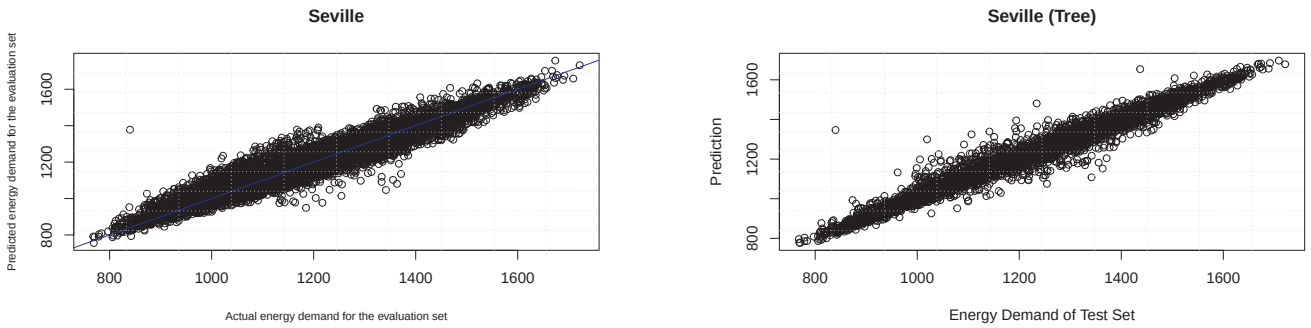

Figure 15: Predicted and Actual Energy Demand Comparison (MLR left, Tree right)

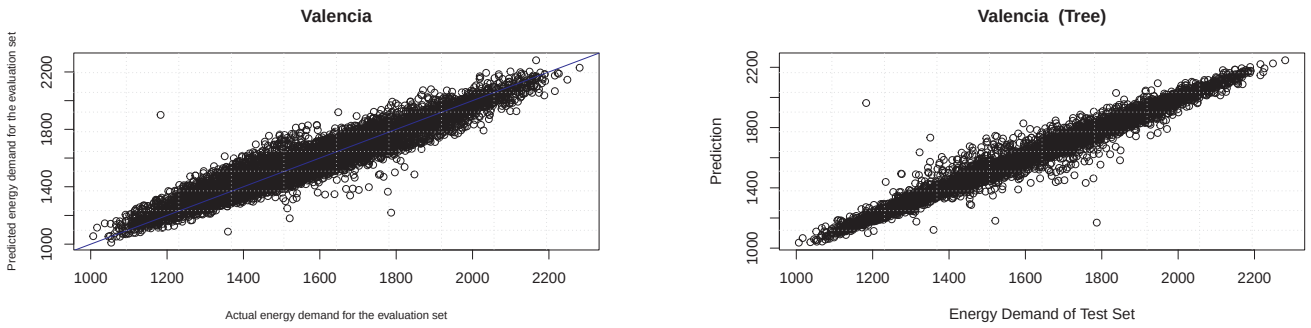

Figure 16: Predicted and Actual Energy Demand Comparison (MLR left, Tree right)

Predictor Importance graph from Tree Based Model

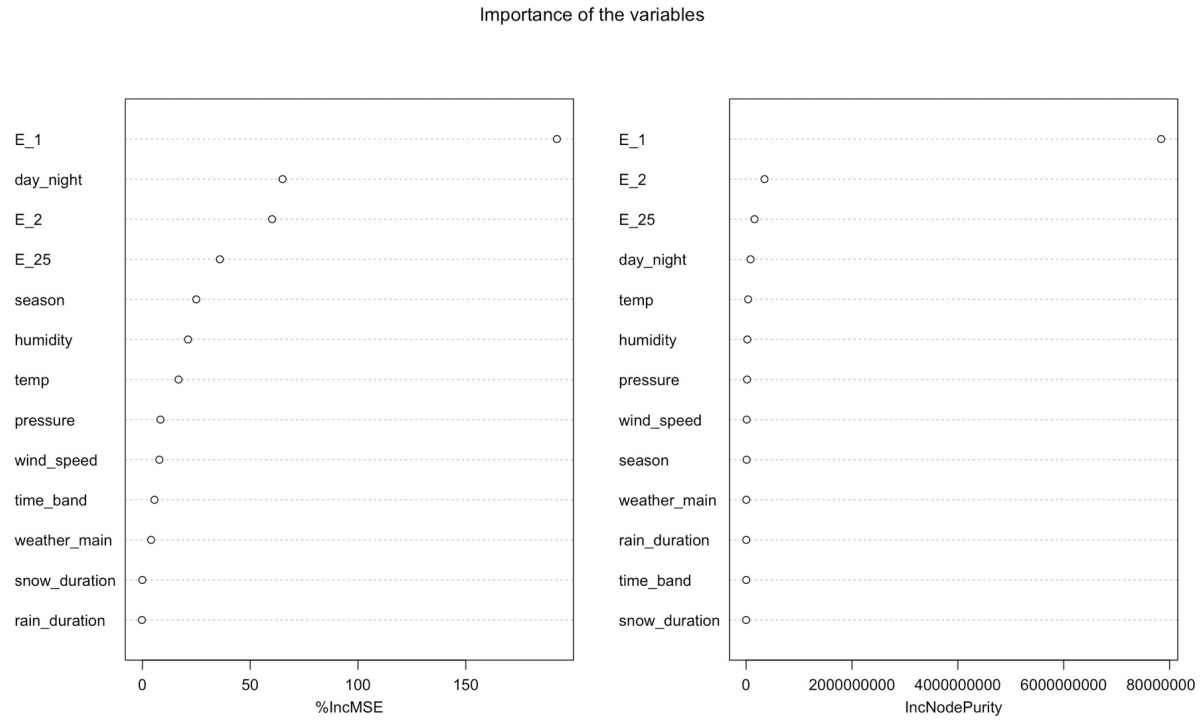

Figure 17: Importance of Predictors in rF BCN

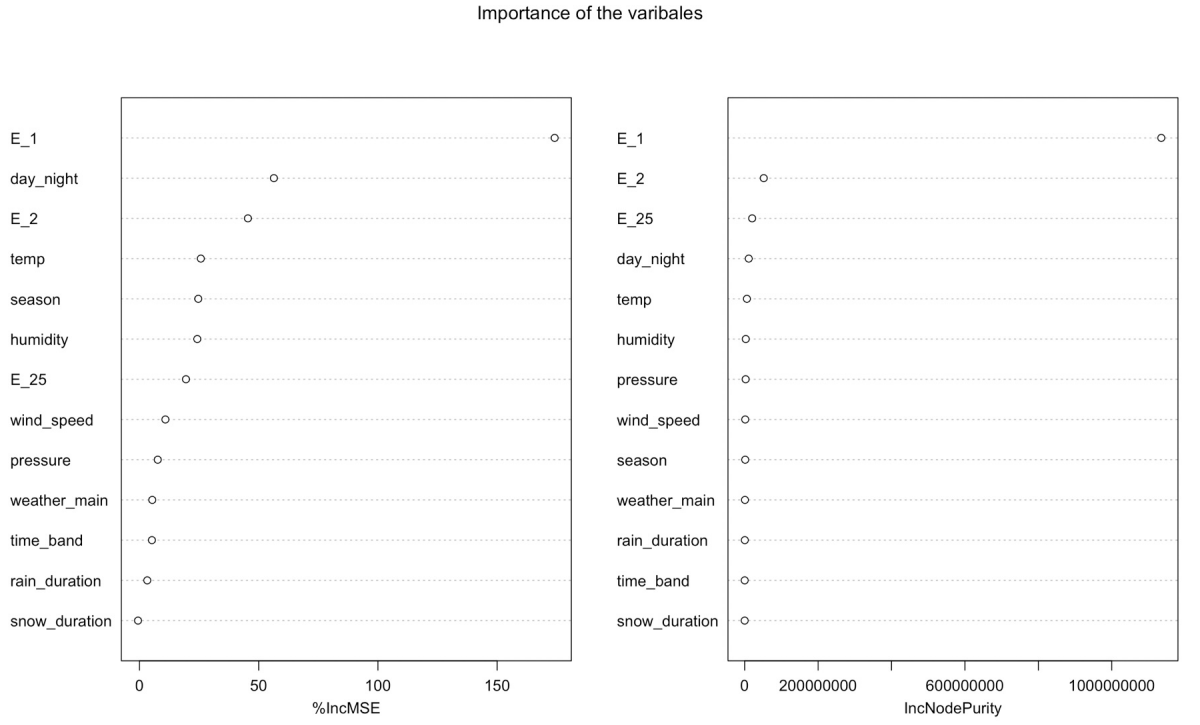

Figure 18: Importance of Predictors in rF BLB

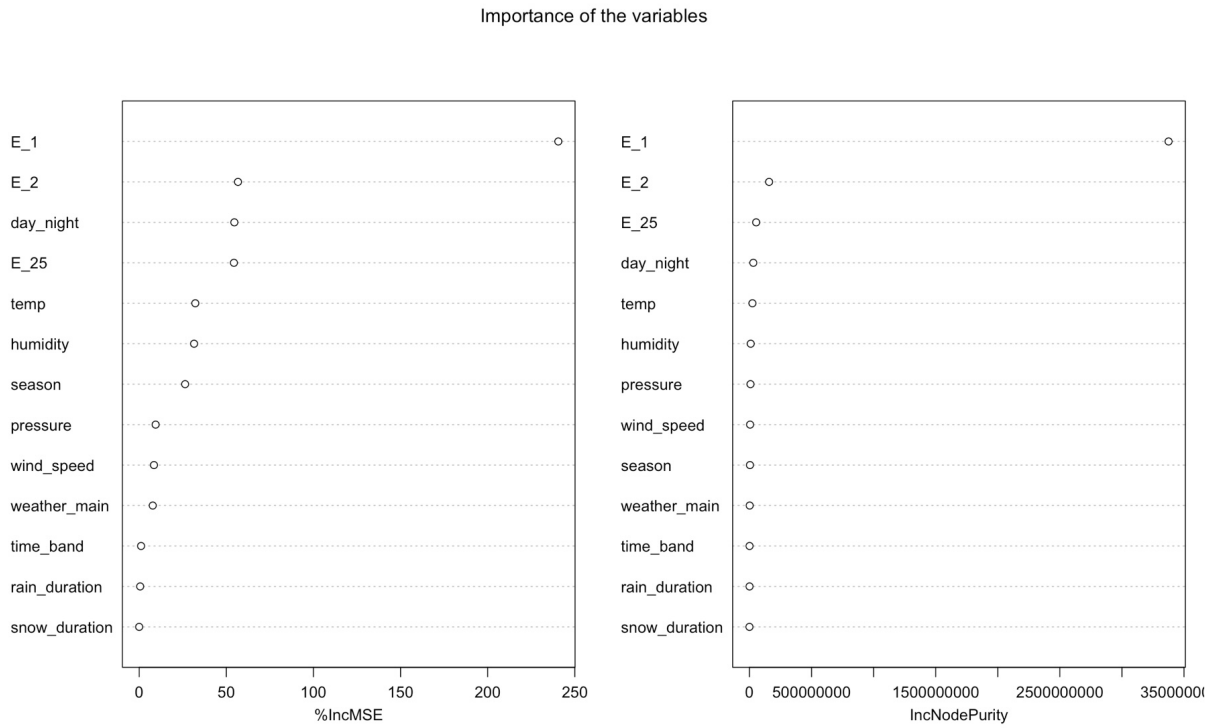

Figure 19: Importance of Predictors in rF MDD

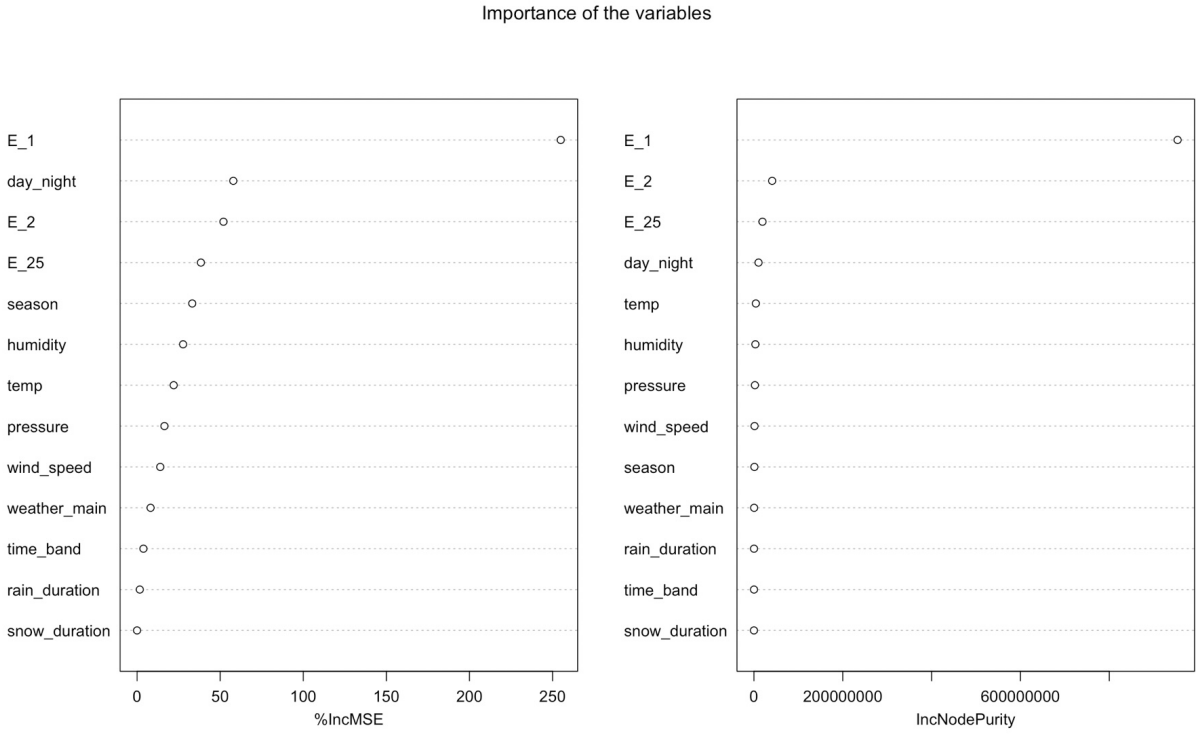

Figure 20: Importance of Predictors in rF SVL

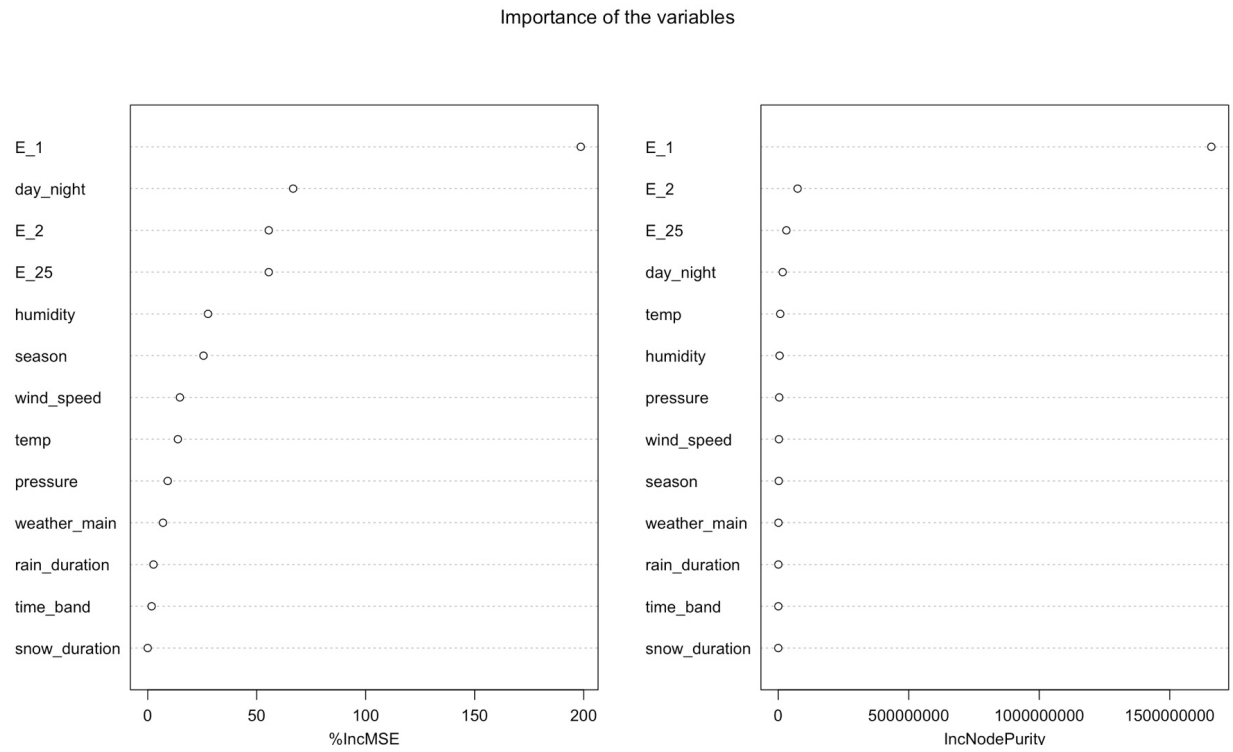

Figure 21: Importance of Predictors in rF VCL

Weather Data

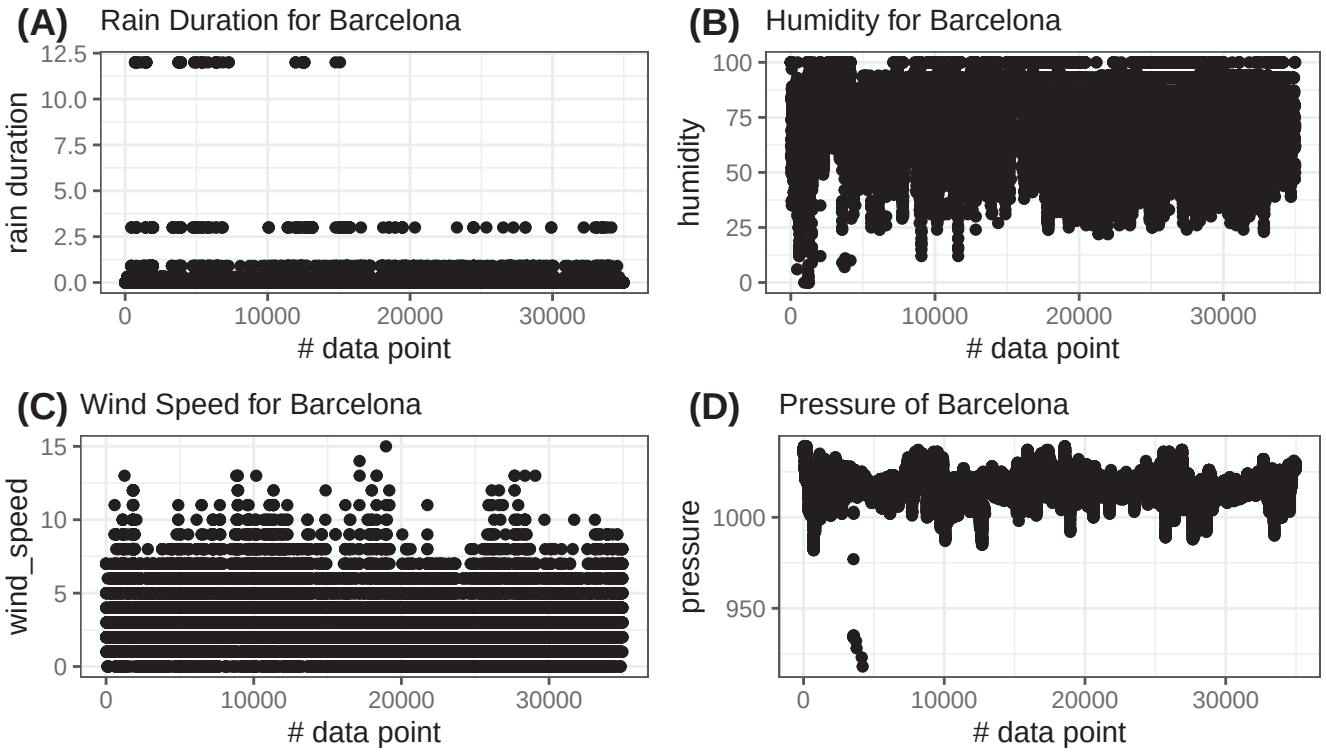

Figure 22: Barcelona Weather Data

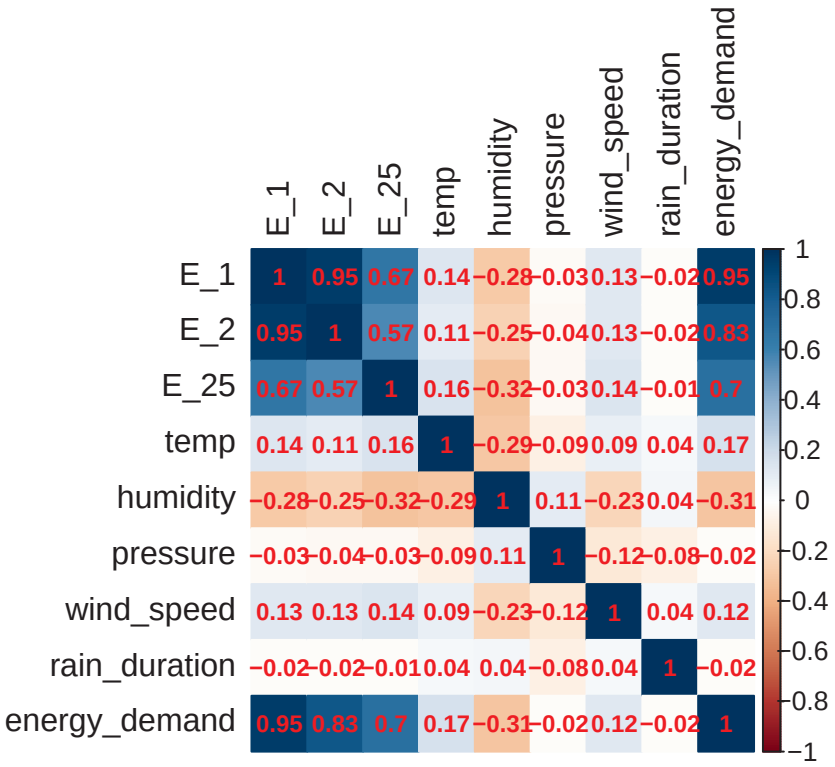

Figure 23: Correlation Spectrum for Barcelona

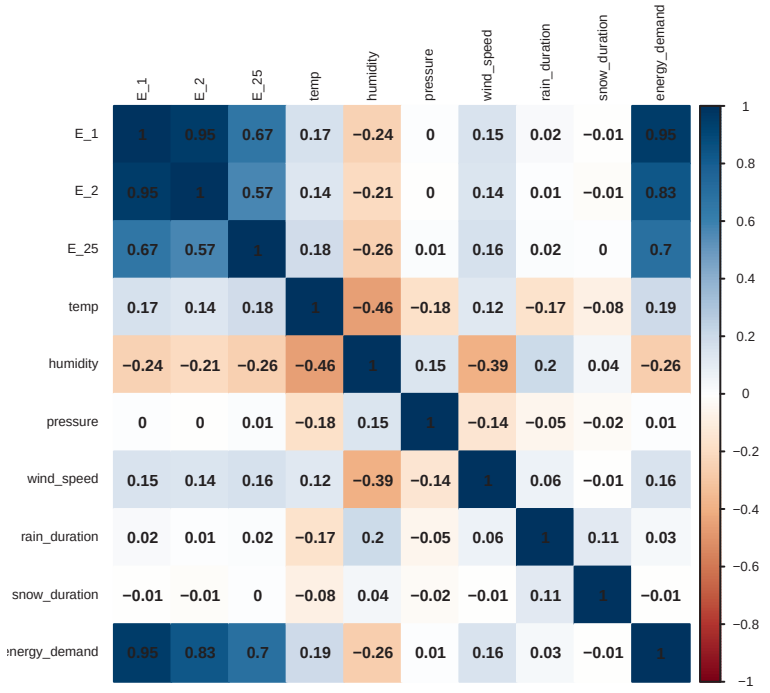

Figure 24: Correlation Spectrum for Bilbao

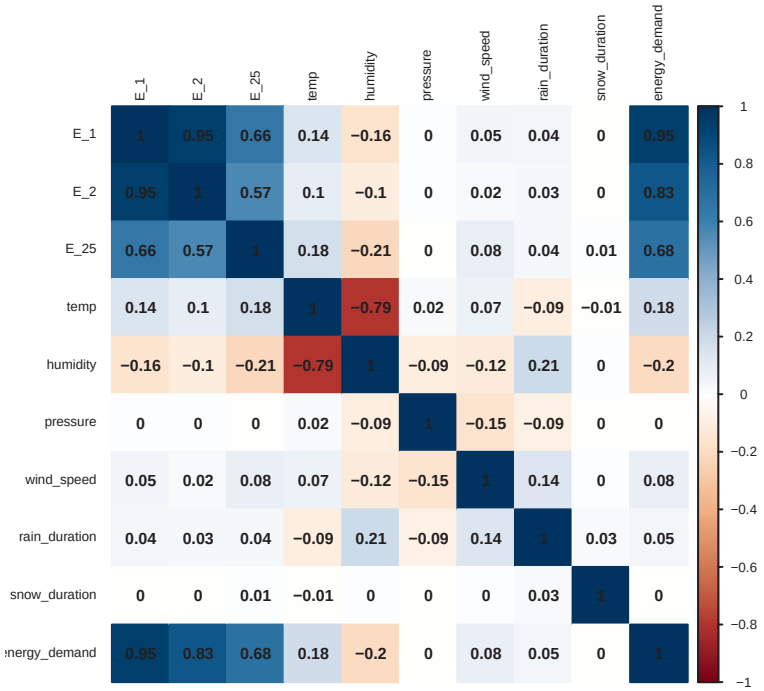

Figure 25: Correlation Spectrum for Madrid

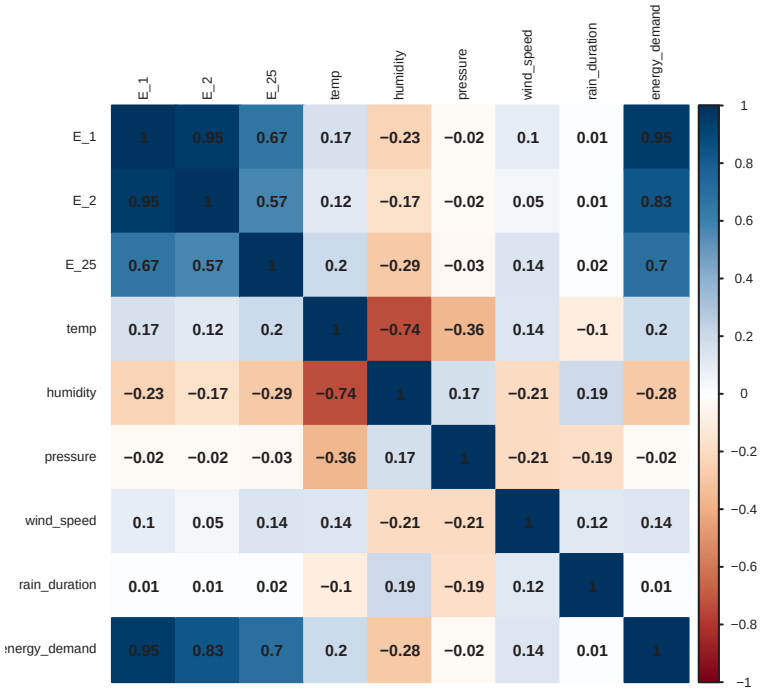

Figure 26: Correlation Spectrum for Seville

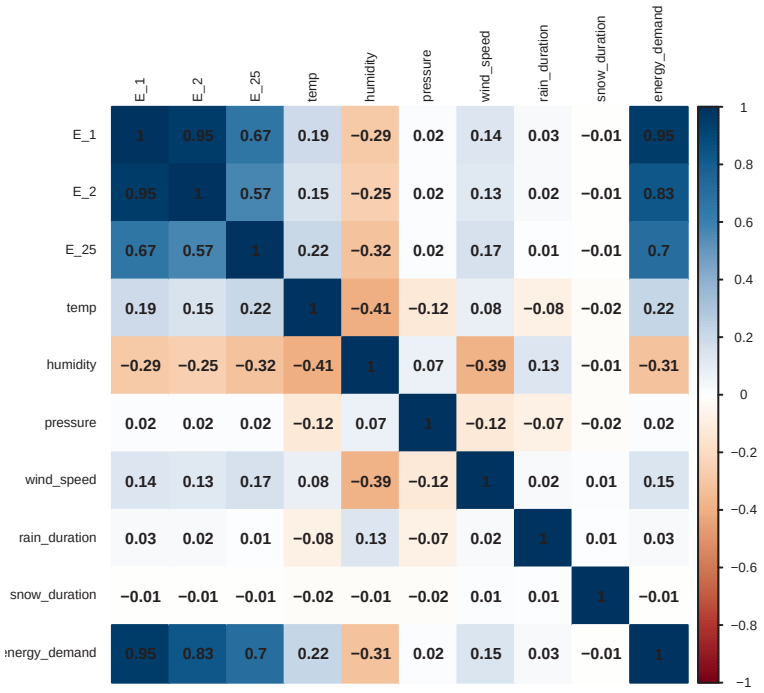

Figure 27: Correlation Spectrum for Valencia

Modeling approach	Variables	Coefficients	Adjusted R^2
Full	13	23	0.9153865
P-Value	12	22	0.9153886
Bestfits	9	16	0.9148913
AIC	12	22	0.9153886
Cross-validation	11	21	0.9153680
OLS-step	12	22	0.9153886

Table 6: Bilbao Reduced Model Analysis Summary

Table 1 illustrates the ANOVA omodel comparison, where the bottom model in each step pairing is the larger (more variable) construct. Progression occurs left to right. For instance, the first comparison was the P-Value reduced model vs the Full model, with the P-Value reduced model being preferred (ANOVA comparison $p - value > 0.05$ (5% significance), meaning H_0 is rejected and no statistical evidence supports the larger model being preferred over the reduced model). This comparison proceeded with each of the successive model types (no comparison was run if the model reduction analysis yielded the same model construct - i.e., variable listing - as that of the P-Value model) and in each step, the P-Value model came out as the preferred model. Furthermore, the following comparison tables provide insight into the Adjusted R-squared value for each of the models analyzed. In all cases, focus is on the model with the combination of highest Adjusted R-squared and fewest corresponding variables / parameters.

Modeling approach	Variables	Coefficients	Adjusted R^2
Full	13	23	0.9160325
P-Value	12	22	0.9160309
Bestfits	8	11	0.9153598
AIC	12	22	0.9160309
Cross-validation	11	21	0.9159974
OLS-step	12	22	0.9160309

Table 7: Madrid Reduced Model Analysis Summary

Modeling approach	Variables	Coefficients	Adjusted R^2
Full	12	24	0.9199264
P-Value	11	23	0.9199208
Bestfits	8	10	0.9189457
AIC	12	24	0.9199264
Cross-validation	11	23	0.9199208

Table 8: Seville Reduced Model Analysis Summary

Modeling approach	Variables	Coefficients	Adjusted R^2
Full	12	24	0.9199264
P-Value	11	23	0.9199208
Bestfits	8	10	0.9189457
AIC	12	24	0.9199264
Cross-validation	11	23	0.9199208

Table 9: Valencia Reduced Model Analysis Summary

Coefficient	Description	Type	(Sub)levels if Categorical
β_0	Intercept		
β_1	E-1, temporal variable for actual hourly energy demand one hour prior	Numeric	
β_2	E-2, temporal variable for actual hourly energy demand two hours prior	Numeric	
β_3	E-25, temporal variable for actual hourly energy energy demand 25 hours prior	Numeric	
β_4	temp, ambient temperature (in Kelvin)	Numeric	
β_5	humidity (%)	Numeric	
β_6	ambient pressure (in hectropascals)	Numeric	
β_7	wind speed (meters/second)	Numeric	
β_8	rain duration (accumulation in millimeters)	Numeric	
β_9	snow duration (accumulation in millimeters)	Numeric	
β_{10}	day or night	Categorical	1 (nighttime; default is daytime)
β_{11}	pricing time band	Categorical	2 (mid-peak and peak; default is offpeak)
β_{12}	season	Categorical	3 (in order: summer, autumn and winter; default is spring)
β_{13}	weather main (description)	Categorical	11 (in order: clouds, drizzle, dust, fog, haze, mist, rain, smoke, snow, squall and thunderstorm; default is clear)

Table 10: MLR model variable designation and description