ABAV Assignment C

by ISMAIL ESACK DAWOODJEE .

Submission date: 20-Sep-2020 10:07PM (UTC+0800)
Submission ID: 1390410520
File name: 32516_ISMAIL_ESACK_DAWOODJEE_._ABAV_Assignment_C_20811_410377810.pdf (2.2M)
Word count: 6735
Character count: 37572

ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT045-3-M-ABAV

ADVANCED BUSINESS ANALYTICS AND VISUALIZATION

APUMF1908DSBA

HAND OUT DATE: 04 SEPTEMBER 2020

HAND IN DATE: 21 SEPTEMBER 2020

WEIGHTAGE: 50%

A Story of Data and Energy: Renewable Energy Generation in Spain

NAME:	Ismail Dawoodjee
TP NUMBER:	TP054033
EMAIL:	tp054033@mail.apu.edu.my
LECTURER:	Dr. Preethi Subramanian

A Story of Data and Energy: Renewable Energy Generation in Spain

Ismail Esack Dawoodjee
Asia Pacific University of Technology and Innovation
tp054033@mail.apu.edu.my

Table of Contents

1. Introduction 3
2. Business Case 4
2.1. Business Goals 4
2.2. Aim and Objectives 5
2.3. Scope 5
3. Methodology 6
4. Data Exploration and Preparation 7
4.1. Loading in the Datasets 7
4.2. Exploring the Variables 8
4.3. Data Preprocessing 9
4.4. Time Series Data Preparation 12
4.4.1. Preparation for Exponential Smoothing Forecasts 12
4.4.2. Preparation for Similarity and Cluster Analyses 13
5. Model Construction, Optimization and Validation 14
5.1. Exponential Smoothing Forecast Models 14
5.2. Time Series Similarity and Cluster Analyses 16
6. Critical Interpretation of Outcomes 19
6.1. Interpretation of Similarity and Cluster Analysis 19
6.2. Interpretation of Price and Load Forecasts 26
7. Discussion and Conclusion 28
References 30

1. Introduction

According to the Intergovernmental Panel on Climate Change (IPCC), the electric power industry has been estimated to contribute around 25% of annual human-caused greenhouse gas emissions (IPCC, 2014). On the other hand, this industry also generates large amounts of data, making it an ideal sector for the application of artificial intelligence (AI) and machine learning (ML) techniques (Rolnick et al., 2019). Hence, ML has enormous potential to transform the electric power industry into a greener and more efficient system. This includes accelerating the research and development of clean energy sources (such as hydro, solar and wind power generation), improvements in energy demand forecasts, optimizing the supply and distribution of electricity with data-driven "smart grids", and reducing the wastage of electrical energy during transportation (Rolnick et al., 2019).

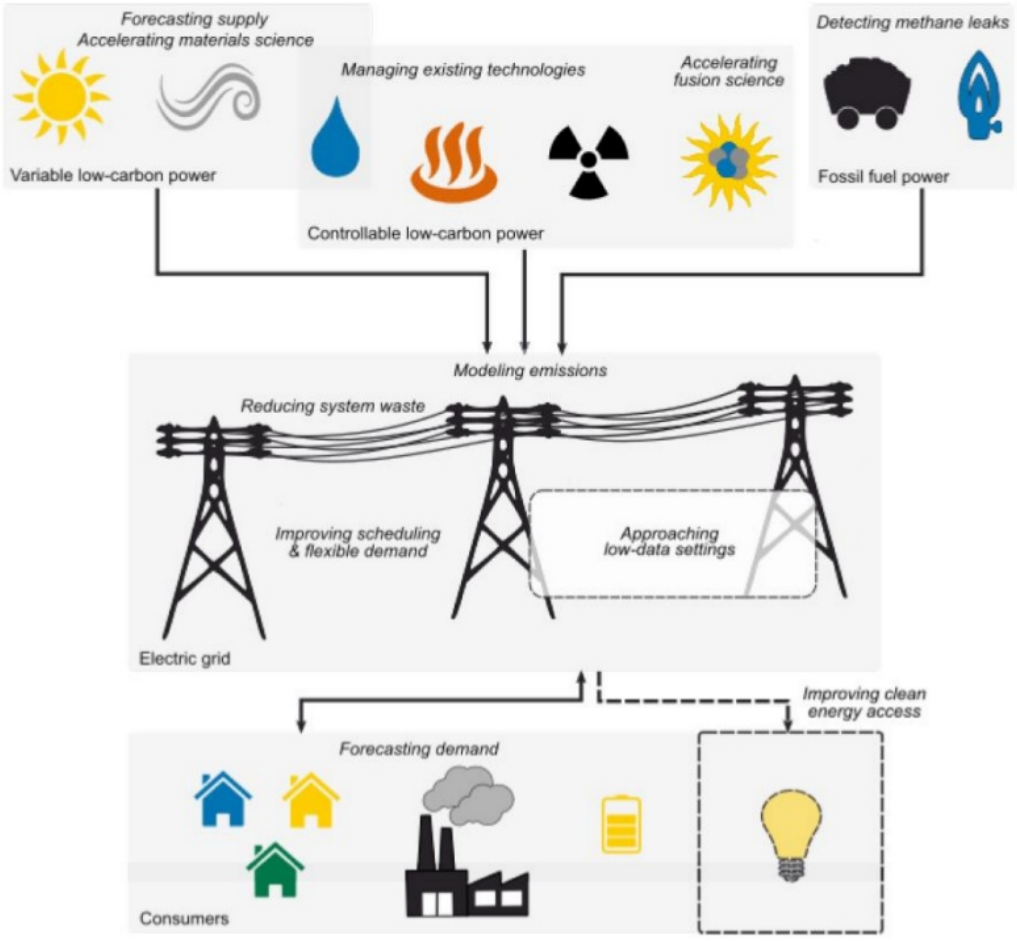


Figure 1: The variety of ways in which machine learning can assist the electric power industry in reducing greenhouse gas emissions (Rolnick et al., 2019).

2. Business Case

For this particular case study, the electric power industry in question comes from Spain, a country well-known for its sunny skies and beautiful coastal beaches. The energy service company, named as Red Eléctrica de España (REE), has gathered four years of electrical demand and price data (from January 2015 to December 2018), as well as the various types of non-renewable (coal, oil, gas, nuclear) and renewable energy generation data.

Figure 2: The land of sun and blue skies (Solucion Asesores XXI, 2016).

2.1. Business Goals

Currently, the company is looking to expand its renewables sector but has not yet decided on which energy source or where the energy generator (e.g. solar farm, wind farm, tidal power station) should be located. To facilitate this decision, weather data from the five largest cities in Spain were extracted from the Open Weather API over the same four year period (OpenWeather Ltd., 2020). This data contains information on each city's temperature, wind speed and direction, rainfall and cloudiness, among others. By analyzing both datasets, REE will be able to make an informed and data-driven decision, not only to figure out the best location to establish a clean energy generator but also to discover how weather patterns influence electricity demand and energy generation.

Justification: Investing in renewable energy resources is an essential stepping-stone to reducing greenhouse gas emissions and slowing down climate change (Rolnick et al., 2019). Spain, as a country surrounded by water on three sides, has already been subject to the ravages of climate

change. The average temperature has risen by 8 °C, and overall precipitation received across the country is 25% less than that of 50 years ago (Heggie, 2020). In addition, about 90% of ice glaciers in the Pyrénées (a mountain range separating Spain from France) has completely disappeared (Heggie, 2020). Understandably, Spain has made and is still making outstanding efforts to combat climate change, and REE is contributing to this cause.

2.2. Aim and Objectives

The main aim of REE is to identify the optimum location (out of the five cities) in which a sustainable energy source can be placed to maximize electricity generation. For example, the city of Seville, which is located in the southern part of Spain, enjoys more sunshine than Bilbao, which is located in the northern mountainous region. Hence, constructing a solar farm near Seville could be *one* reasonable choice. Given the above aim, the objectives are:

1. Descriptive Analytics: To analyze, describe, and visualize trends in the energy generation sources, consumer energy demand, electricity price, and fluctuations in weather.
2. Predictive Analytics: To build multiple linear regression models to forecast consumer demand and energy generation, based on the city and weather predictor variables.
3. Prescriptive Analytics: To anticipate and prepare for future surges in demand, reduce the occurrence of surpluses and shortages in electricity, and put forward impactful recommendations for the location of several possible renewable energy generators.

2.3. Scope

The datasets that are used in this study can be obtained directly from Kaggle[1] (Jhana, 2019). Alternatively, the energy generation data can be manually extracted from the ENTSO-E (European Network of Transmission System Operators for Electricity) website (ENTSO-E, 2020), price data from the REE company website (REE, 2020), and the weather dataset can be purchased from the OpenWeather API page (OpenWeather Ltd., 2020).

The scope of this study will be limited to analyzing only two datasets: the energy and weather datasets directly obtained from the Kaggle page. No external datasets will be used, nor will there be an analysis of data that is outside the time period from January 2015 to December 2018. Moreover, the analysis will be limited only to business- and data-related problems that can be solved *before* the renewable energy generator construction takes place. Problems that occur during and after the construction takes place (that may still be solved with ML and AI), such as the logistics of transporting electricity, or the physics and engineering cross-disciplinary issues related with optimizing the generation of electricity, will not be considered. Finally, only weather

[1] https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather

data from the five largest Spanish cities (namely, Madrid, Bilbao, Seville, Barcelona, and Valencia) will be analyzed. Other cities or regions in Spain will not be considered.

3. Methodology

The data mining approach to achieve the aim and objectives described above will be the CRISP-DM (CRoss Industry Standard Process for Data Mining) analytics model. This model includes the vital first step of understanding the business and setting goals for it to accomplish. Henceforth, all subsequent steps and objectives in the case study can be meaningfully retraced back to the original goal that REE has set out to achieve: to expand its renewables sector in the name of slowing down climate change, but also making a good profit from it.

For the first objective, the visualization and data mining tools Tableau and SAS Enterprise Miner will be used for preliminary data exploration, preprocessing and feature engineering. The trend over the years can be visualized through the use of time series line plots, with time aggregation (resampling by day, week or month) and variable aggregation (aggregating over all renewable or non-renewable energy sources). An interactive dashboard linking together some of the most important dependencies between city weather, energy, demand, and price would be an excellent visualization of the hidden patterns that exist within the two datasets.

Next, linear regression forecasting models can be used to forecast for future energy demand and generation, which will be taken as the dependent variables. The reason for using a regression model is because sustainable energy generation is heavily dependent on the weather (e.g. the sun needs to be shining on the photovoltaic cells to generate energy and the wind needs to be blowing at a certain speed for the wind turbines to rotate and produce electricity). Hence, weather variables can be inserted into the regression model as the independent predictor variables, after accounting for their predictive power[2] using the adjusted R^2 value, Akaike's Information Criterion, and other measures of predictive accuracy (Hyndman and Athanasopoulos, 2018).

In addition, knowledge about how consumer energy demand varies with weather variables can be obtained from a similar regression model. For instance, people tend to use heaters on cold days and air conditioners on hot days, but what about on windy or cloudy days? Evidently, the effect of weather on human behavior may not be as predictable as that on mechanical objects. Nevertheless, the regression models will make a brave attempt at modelling both energy demand and energy generation.

Finally, based on the regression models, forecasts can be made for the next 24 hours, 7 days, 4 weeks and 3 months. Although it is an inevitable fact that forecasts become less accurate as the forecast horizon gets larger, the overall trend and seasonality data obtained from the

[2] Predictive power: how important an independent variable is in influencing the dependent variable.

predictions will help REE make better decisions in the future. In other words, the company's response to consumer demand can be more effective and less wasteful due to closing the occasional gap between energy supply and demand.

4. Data Exploration and Preparation

4.1. Loading in the Datasets

A new project named *ABAV Assignment C* was created on SAS Enterprise Miner (EM), along with a new diagram with the same name. The two datasets introduced in Section 2.1 are on the user's system and not on the SAS library, so they will need to be loaded first. Loading an external dataset can be done by using the **File Import** node on the **Sample** tab of the EM toolbar, as shown in Figure 3 below:

Figure 3: The File Import node on the Sample tab of the EM toolbar

Since there are two datasets, two **File Import** nodes are required. After placing the two nodes on the diagram, the **Advanced Advisor** property of both nodes were changed from "No" to "Yes". The datasets can then be imported normally from the **Import File** property of each node. Finally, the nodes were **Ran** and renamed to *Energy Data* and *Weather Data* to reflect the corresponding imported files, shown in Figure 4 below:

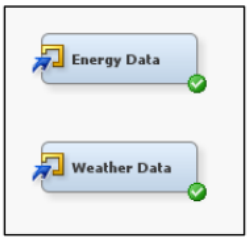


Figure 4: The two imported csv files on the named File Import nodes

One difficulty faced while importing the energy dataset was that the header row contains variable names with spaces in between. For example, *generation fossil oil*. This caused an error in which EM found 0 observations in the dataset, even though this was not true. The difficulty was overcame by replacing all spaces in the header row with an underscore, so that *generation fossil oil* now becomes *generation_fossil_oil*, which promptly resolved the error. Moreover, two variables had very long names, so these were changed to be slightly shorter. Finally, the variable *dt_iso* signifying time variable in the *Weather Data*, was changed to *time* to match the time variable name in the *Energy Data*.

4.2. Exploring the Variables

Variables in each dataset can be explored by right-clicking the **File Import** node and choosing **Edit Variables**, or by going to the **Variables** property of the node. In the energy dataset, there are 29 variables and 35064 observations (one for each hour of the 4 years from 2015 to 2018), seen by going to the **Results** page of the node. Out of the 29 variables, there were 2 variables with 100% missing entries and 6 variables that contain only 0's, referred to as "Unary" variables by EM. These 8 variables were rejected and dropped from the energy dataset.

In addition, the remaining 21 variables consist of one time variable, various renewable and non-renewable energy generation data, and the day-ahead price, total load, wind and solar energy forecasts as well as their actual values. These four variables were rejected because they are company forecasts, and they should not contribute as predictor variables in the process flow. The *price_actual* and *total_load_actual* variables were set to the **Target** role because they are the actual dependent variables that will be forecasted using the other input variables. The missing percentages are very small, ranging from 0% to 0.1%, with many of the variables having around 0.05% missing entries. All the energy dataset variables are shown in Figure 5 below:

Name	Role	Level	Report	Order	Drop	Percent Missing
generation_other	Input	Interval	No		No	0.051335
generation_nuclear	Input	Interval	No		No	0.048483
generation_solar	Input	Interval	No		No	0.051335
generation_other_renewable	Input	Interval	No		No	0.051335
generation_hydro_run_of_river_an	Input	Interval	No		No	.
generation_hydro_pumped_consumpt	Input	Interval	No		No	0.054187
generation_hydro_water_reservoir	Input	Interval	No		No	0.051335
time	Time ID	Interval	No		No	0
price_day_ahead	Rejected	Interval	No		No	0
total_load_forecast	Rejected	Interval	No		No	0
total_load_actual	Target	Interval	No		No	0.102669
generation_wind_onshore	Input	Interval	No		No	0.051335
generation_waste	Input	Interval	No		No	0.054187
price_actual	Target	Interval	No		No	0
generation_fossil_gas	Input	Interval	No		No	0.051335
generation_fossil_brown_coal_lig	Input	Interval	No		No	.
generation_fossil_hard_coal	Input	Interval	No		No	0.051335
forecast_wind_onshore_day_ahead	Rejected	Interval	No		No	0
forecast_solar_day_ahead	Rejected	Interval	No		No	0
generation_biomass	Input	Interval	No		No	0.054187
generation_fossil_oil	Input	Interval	No		No	0.054187
generation_geothermal	Rejected	Unary	No		Yes	0.051335
generation_marine	Rejected	Unary	No		Yes	0.054187
generation_hydro_pumped_aggregat	Rejected	Unary	No		Yes	100
forecast_wind_offshore_eday_ahea	Rejected	Unary	No		Yes	100
generation_wind_offshore	Rejected	Unary	No		Yes	0.051335
generation_fossil_oil_shale	Rejected	Unary	No		Yes	0.051335
generation_fossil_peat	Rejected	Unary	No		Yes	0.051335
generation_fossil_coal_derived_g	Rejected	Unary	No		Yes	.

Figure 5: The 17 accepted variables, 12 rejected unary and pre-forecast variables in the energy dataset

For the weather dataset, there are 17 variables and 178396 observations, with no missing entries. However, four of the variables (namely, *weather_main*, *weather_icon*, *weather_id* and *weather_description*) convey the same information conceptually, so only *weather_main* was kept and the other three were dropped. In addition, *weather_icon* and *weather_description* had over 20 categories, so EM automatically rejected them. All the weather dataset variables are shown in Figure 6 below:

Name	Role	Level	Report	Order	Drop	Percent Missing
temp_max	Input	Interval	No		No	0
temp_min	Input	Interval	No		No	0
snow_3h	Input	Interval	No		No	0
temp	Input	Interval	No		No	0
wind_deg	Input	Interval	No		No	0
wind_speed	Input	Interval	No		No	0
weather_main	Input	Nominal	No		No	0
city_name	Input	Nominal	No		No	0
humidity	Input	Interval	No		No	0
dt_iso	Time ID	Interval	No		No	0
clouds_all	Input	Interval	No		No	0
rain_1h	Input	Interval	No		No	0
rain_3h	Input	Interval	No		No	0
pressure	Input	Interval	No		No	0
weather_icon	Rejected	Nominal	No		Yes	0
weather_description	Rejected	Nominal	No		Yes	0
weather_id	Rejected	Interval	No		Yes	0

Figure 6: The 14 accepted variables and 3 rejected categorical weather variables

4.3. Data Preprocessing

Firstly, the datasets are already in the format required by EM to be considered a times series data. Both contain a single Time ID variable with an hourly frequency, and all other variables (excluding the rejected ones) are in the interval scale except for *weather_main* and *city_name*. These two nominal variables were changed to Cross ID variables, which are non-interval variables that are used for time series data aggregation purposes. The final variable table is shown in Figure 7 below:

Name	Role	Level	Report	Order	Drop
temp_min	Input	Interval	No		No
time	Time ID	Interval	No		No
temp	Input	Interval	No		No
temp_max	Input	Interval	No		No
wind_deg	Input	Interval	No		No
wind_speed	Input	Interval	No		No
weather_main	Cross ID	Nominal	No		No
city_name	Cross ID	Nominal	No		No
pressure	Input	Interval	No		No
humidity	Input	Interval	No		No
clouds_all	Input	Interval	No		No
rain_3h	Input	Interval	No		No
snow_3h	Input	Interval	No		No
rain_1h	Input	Interval	No		No
weather_icon	Rejected	Nominal	No		Yes
weather_descrip	Rejected	Nominal	No		Yes
weather_id	Rejected	Interval	No		Yes

Figure 7: The updated weather variable roles to be used for time series data preparation

Secondly, the weather data contains "duplicate" observations. It contains 178396 observations instead of $175320 = 35064 \times 5$ observations. They are not true duplicates because more than one weather description can be attributed to each hour. However, since the weather description variable was dropped, some of the hours with more than one weather description become exactly the same observation, i.e. they become duplicates. This was dealt with by using a **SAS Code** node from the **Utility** tab of the EM toolbar and specifying the `nodupkey` code while sorting data by city name and time, shown in Figure 8 below. The code was entered by going to the **Code Editor** property of the node. The non-duplicate weather data was then verified to have 175320 observations using the **DMDB** and **StatExplore** nodes.

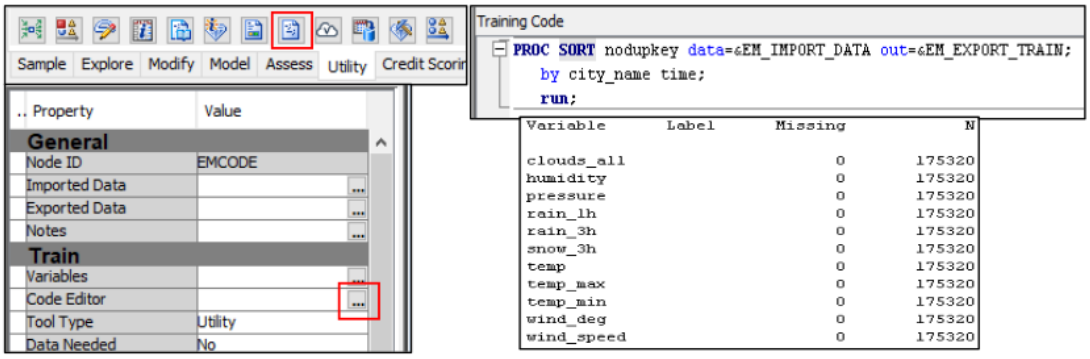


Figure 8: Dropping duplicates using the **SAS Code** node and verification of result

Next, the dataset was filtered by five different **Filter** nodes, where each filter allows observations from each of the five cities. In the property options, the **Default Filtering Method** for both class and interval variables should be set to **None**. Then the city filter can be done by going to **Class Variables**, and applying the filter by removing the undesired categories. For example, in Figure 9 below, only the Barcelona class is needed so the other four cities are shaded out.

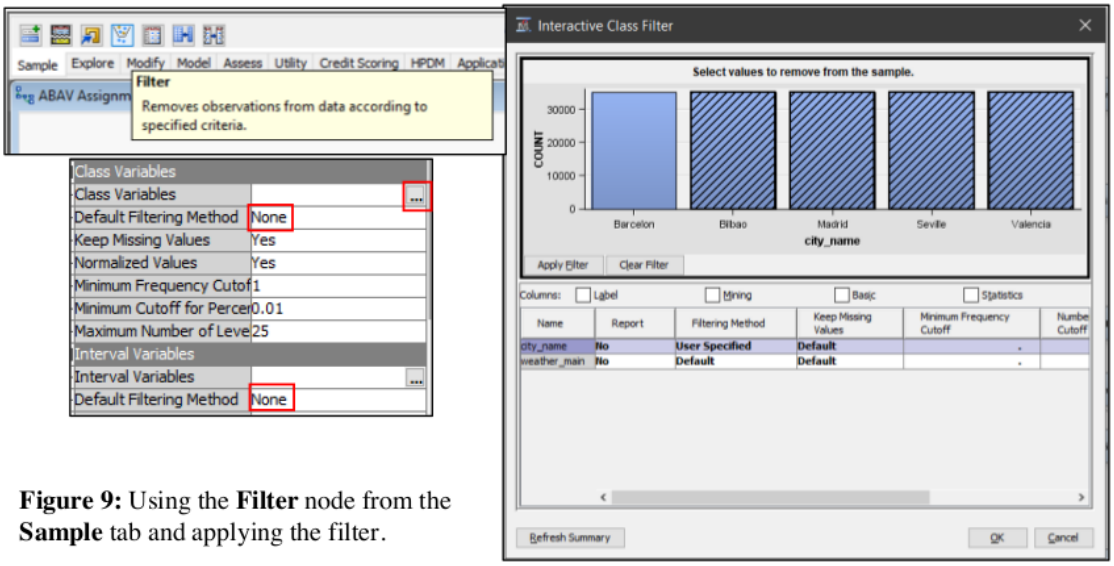


Figure 9: Using the **Filter** node from the **Sample** tab and applying the filter.

Finally, the energy dataset and five city-specific weather datasets can be merged. This was done using the **Merge** node from the **Sample** tab. The merging should be done **By** the *time* variable, which was specified by going to the **Variables** property of the node.

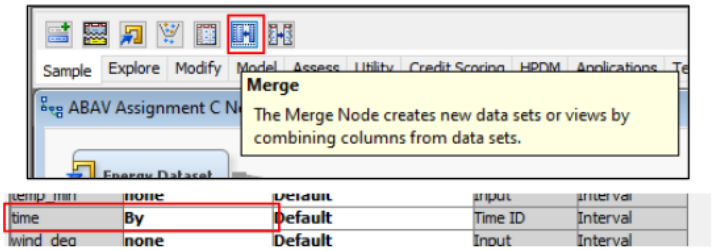


Figure 10: Merging two datasets on the *time* variable

The flow process after the data loading and preprocessing stage is shown in Figure 11 below, resulting in five different datasets. These were also checked using **DMDB** and **StatExplore** nodes.

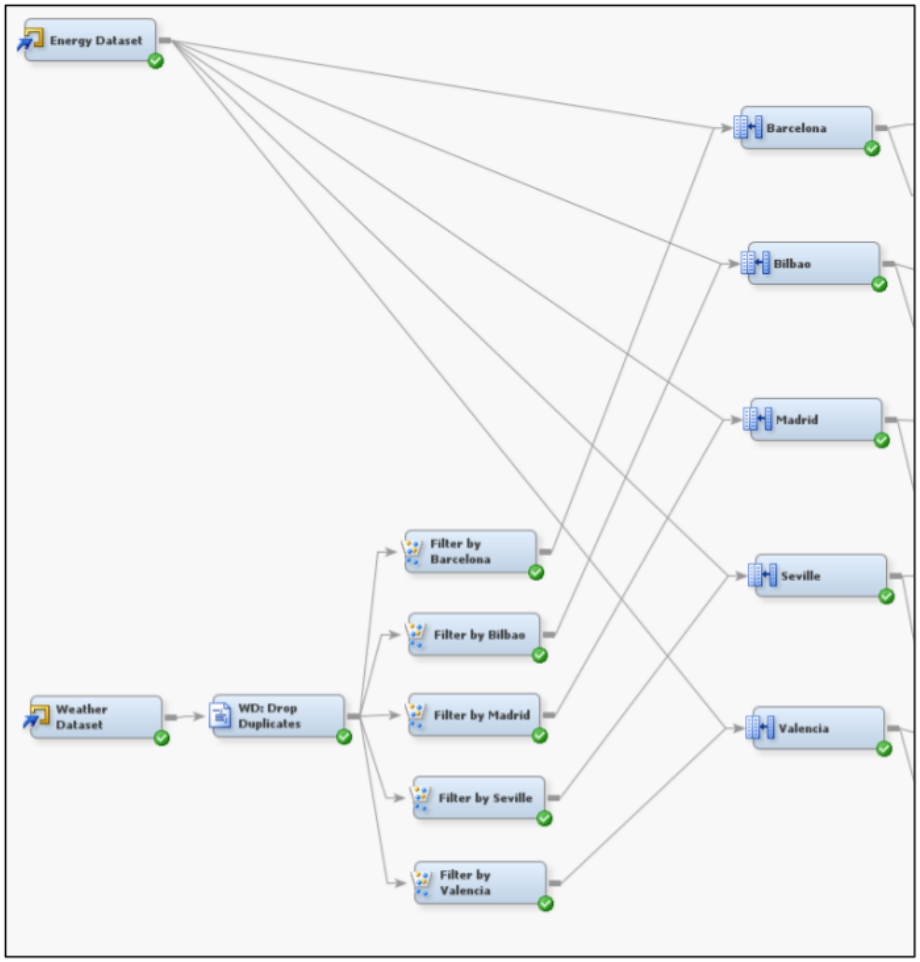


Figure 11: Merging two datasets on the *time* variable

4.4. Time Series Data Preparation

The first objective in Section 2.2 was achieved by exploring the datasets on Tableau. Now, the second objective will be revised as follows:

2. Predictive Analytics: To use exponential smoothing models to forecast consumer energy demand and electricity price.

The reason for using exponential smoothing instead is because of the way the data is already prepared as a time series dataset with a convenient **Time ID**, which makes it natural to use the built-in **TS Data Preparation**, **TS Similarity**, and the **TS Exponential Smoothing** nodes on EM. In addition, the time series similarity analysis could also be used for clustering the weather and energy time series to find out which of them are similar and hence, make a decision on which renewable energy is most suitable for each city.

4.4.1. Preparation for Exponential Smoothing Forecasts

Before carrying out the forecast, the missing data from the energy dataset need to be imputed and the target variables (*price_actual* and *total_load_actual*) need to be aggregated into monthly, weekly or daily data. This task was done using the **TS Data Preparation** node, which provides cleaning and aggregation options, shown in Figure 12 below:

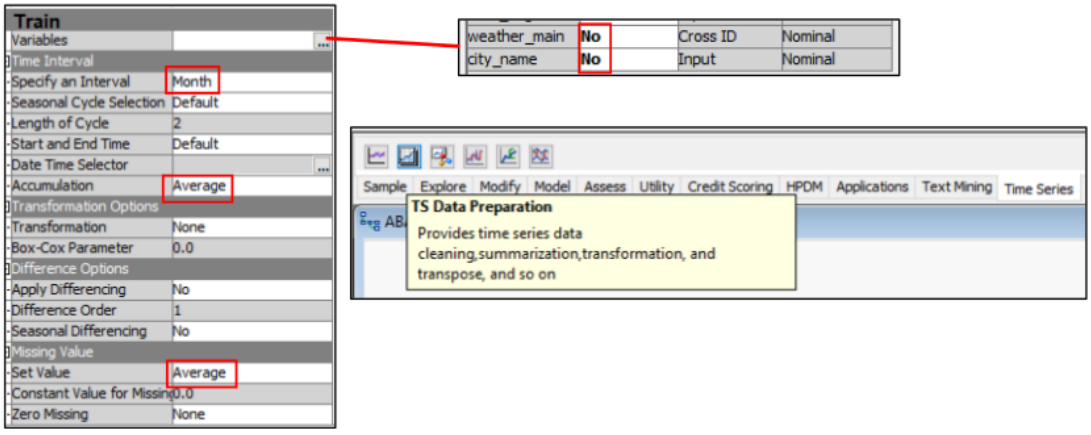


Figure 12: Setting options in the **TS Data Preparation** node

As a starting point, the target variables were aggregated into **Average Monthly** data. Mean imputation was carried out by changing the **Set Value** option to **Average**. Usually, linear interpolation would be preferable, but EM does not provide this option and also, the amount of missing entries are very few so it would not make too much of a difference. The other options can be left as **Default** because specifying the interval is enough to infer the seasonality and EM can automatically detect the start and end times of the time series dataset. Afterwards, when comparing forecast accuracies, **Difference Options** can be changed for experimentation.

4.4.2. Preparation for Similarity and Cluster Analyses

Similarity analysis is used for discovering patterns between different times series and for clustering together the most similar ones. The results from a similarity analysis, in the form of a distance matrix can also be fed as data to perform a cluster analysis. The former is useful for finding out which of the independent variables are most/least similar to a target variable, while the latter (without target variables) is useful for grouping together similar series. Hence, a **Metadata** control node was used first to allow switching between specifying targets and rejecting them. After that, the usual **TS Data Preparation** node can be used for aggregation, imputation and preprocessing for similarity/cluster analysis. The entire data loading and preparation process flow is shown in Figure 13 below:

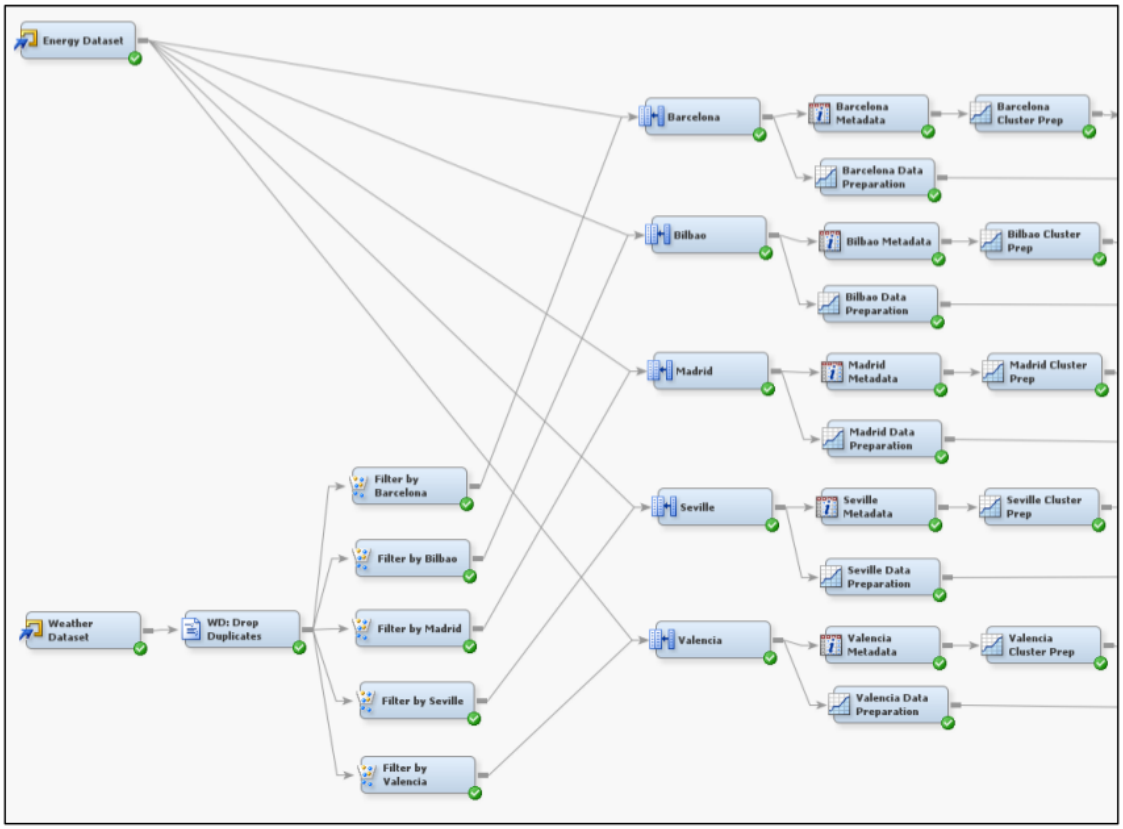


Figure 13: The data loading, filtration, merging and time series data preparation process flow

5. Model Construction, Optimization and Validation

5.1. Exponential Smoothing Forecast Models

For each city-specific dataset, both target variables *price_actual* and *total_load_actual* were forecasted for 6 months (after monthly aggregation), 6 weeks (after weekly aggregation), 7 days (after daily aggregation) and 24 hours (using the original hourly data points). For all forecasts, the **Smooth Outliers** option was set to **Yes** with the replacement set to **Predicted Value**, which automatically takes care of outliers in the data. The outlier smoothing has the advantage of maintaining the seasonality and trend in the data without being overly influenced by it. In addition, the forecasts indicate outlier values in the plot, which has the benefit of not completely throwing away the outliers and allowing the user to study them more closely. The **TS Exponential Smoothing** node and outlier handling is shown in Figure 14 below:

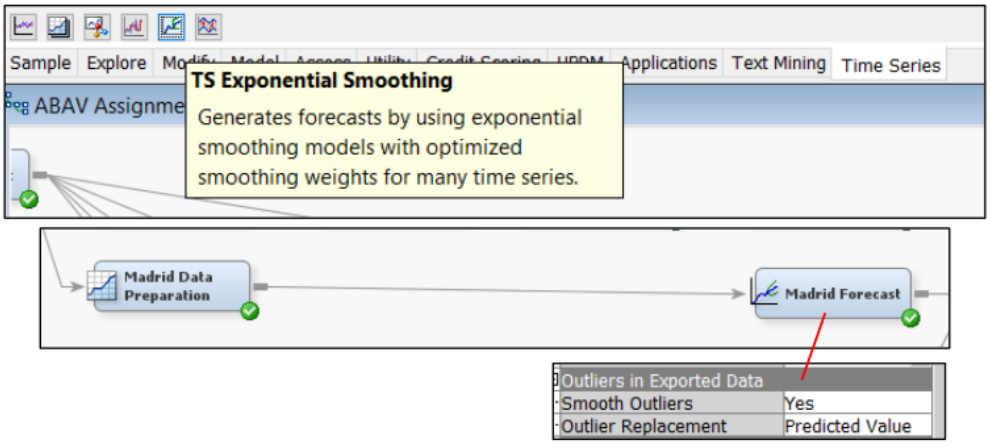
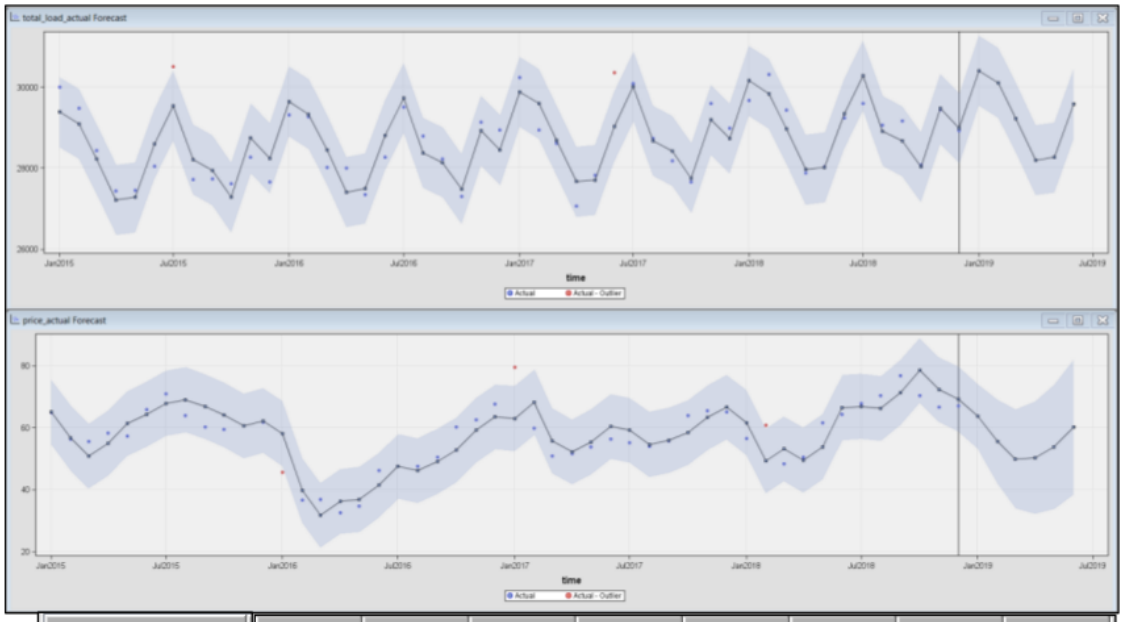


Figure 14: The **TS Exponential Smoothing** node and outlier replacement

A monthly price and demand forecast using the Madrid dataset is shown in Figure 15, along with an evaluation using several metrics. It does not matter which dataset is used because the target and energy variables are the same for all five cities, unless the input weather variables are also being forecasted (which is not the objective here).



Variable Name	Mean Square Error	Root Mean Square Error	Unbiased Mean Square Error	Unbiased Root Mean Square Error	Mean Absolute Percent Error	Mean Absolute Error	R-Square	Adjusted R-Square
price_actual	27.35355	5.230062	28.54284	5.34255	6.978418	3.992842	0.737699	0.731997
total_load_actual	182276.9	426.939	194428.7	440.9407	1.16861	337.7541	0.784764	0.775197

Figure 15: Monthly average forecasts for price and energy load/demand for the next 6 months

The forecast for total load seems to have captured both the seasonality and upward trend of the time series with a good R^2 value of 78% and Mean Absolute Percentage Error, $MAPE = 1.2\%$, while the forecast for price (which does not have a clear seasonality or trend) resulted in a higher error and lower R^2 value. Based on this initial impression, several differencing options were changed. The **Difference Option** in the **TS Data Preparation** node was changed in the following ways, and for each of these options, a forecast was made. The first option created the above plot.

	Apply Diff.	Diff. Order	Load R^2	Load *MAPE*	Price R^2	Price *MAPE*
1	No	(NA)	78.5%	1.17%	73.8%	6.98%
2	No	**Seasonal**	7.27%	184%	69.2%	157%
3	Yes	2	78.9%	435%	47.3%	887000%
4	Yes	3	85.9%	114%	70.8%	127%
5	Yes	4	77.5%	90.9%	74.0%	192%
6	Yes	6	68.3%	97.0%	72.8%	121%
7	Yes	12	7.27%	184%	69.2%	157%

Based on the above table, it is best not to use differencing for the price and load target variables. Differencing may sometimes improve the R^2 value but it comes at the cost of a higher error, similar to the phenomenon of overfitting. Hence, for subsequent forecasts, the differencing option was no longer applied. After experimentation with the differencing option, forecasts using different aggregations were done, and the results are summarized in the table below. The **Forecast Lead** property indicates the forecast horizon for which the forecasting was done. For example, for a semi-month aggregation, the forecast was done for the next 6 semi-months (next 3 months), and for hourly aggregation, the forecast was done for the next 24 hours.

	Aggregation	Forecast lead	Load R^2	Load *MAPE*	Price R^2	Price *MAPE*
1	Hour	24	96.7%	1.99%	96.2%	3.96%
2	Day	7	26.5%	6.19%	84.5%	6.90%
3	Week	6	61.4%	2.37%	81.2%	6.27%
4	Semi-month	6	79.2%	1.47%	80.8%	6.27%
5	Month	6	78.5%	1.17%	73.8%	6.98%
6	Quarter	4	90.3%	0.488%	47.5%	9.69%

5.2. Time Series Similarity and Cluster Analyses

Unlike the price and load forecasts, the cluster analysis should be city-specific in order to group together energy and weather variables. To achieve the third objective of figuring out which city to place a new renewable energy generator (or more generally, the region-specific weather conditions favoring this placement) the similarity between the variable time series were analysed, followed by clustering the similar time series together and segmenting them. If interpretation is not an issue, the Cross ID variable *weather_main* can be used. However, interpreting the many time series that are cross-referenced by *weather_main* categories was found to be difficult, so this variable was Rejected in the **Metadata** node.

Seville Data – The first experiment was done by running similarity and cluster analyses on the Seville dataset. No aggregation was done, so the interval between each observation should be hourly. The number of clusters was chosen to be 7 in both similarity and cluster nodes, as shown in Figure 16 below. The rest of the settings were kept as default.

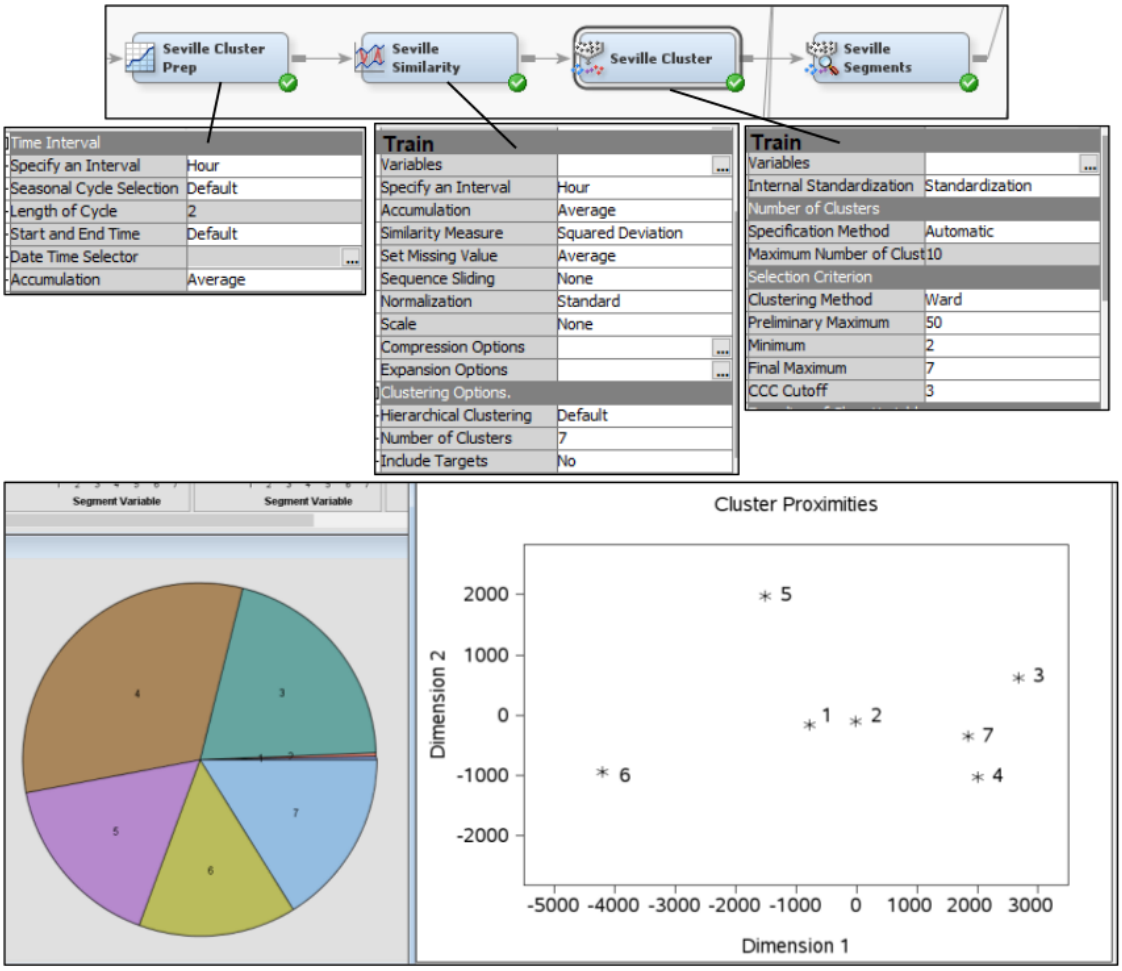
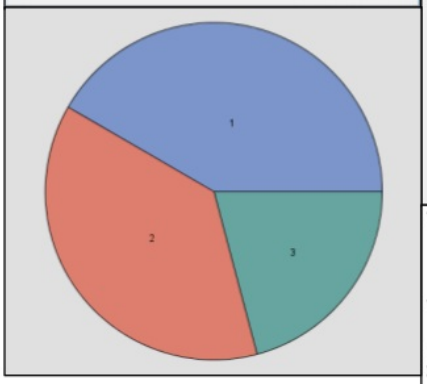
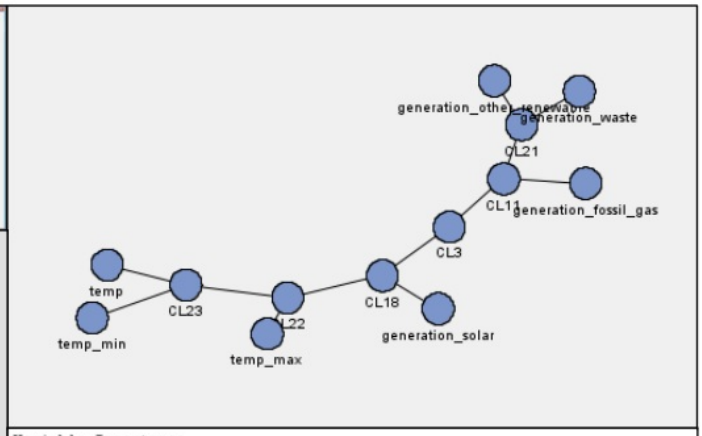


Figure 16: Cluster analysis for the Seville dataset

From the above results, Segments 5 and 6 are separated far from Segments 3, 4 and 7, with Segments 1 and 2 being the outliers. It also indicates that 3 clusters (Segment 5, Segment 6, and Segments 3, 4, 7 together) would be sufficient enough to segment the dataset if the outliers could be taken care of. After this, several experiments were done by varying the maximum number of clusters and changing the aggregation interval. The most promising result with no outliers, clear division and distance between clusters and a sufficient number of important variables in each cluster was obtained with monthly aggregation and three clusters.

Hence, by finally settling on the best result, monthly aggregation was done to smooth out any existing outliers and the option of maximum 3 clusters were chosen. The **Segment Profile** node from the **Assess** tab was also ran, providing the following results:





Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
temp_min		1	0	1.00000
humidity		0	1	0.98953
temp		0	1	0.97895
rain_1h		0	1	0.96825
generation_solar		0	1	0.95743
temp_max		0	1	0.94648
generation_biomass		1	0	0.86716
generation_other_renewable		0	1	0.86716
generation_other		0	1	0.86716
generation_waste		0	1	0.83561
generation_fossil_oil		0	1	0.83561
rain_3h		0	1	0.80283

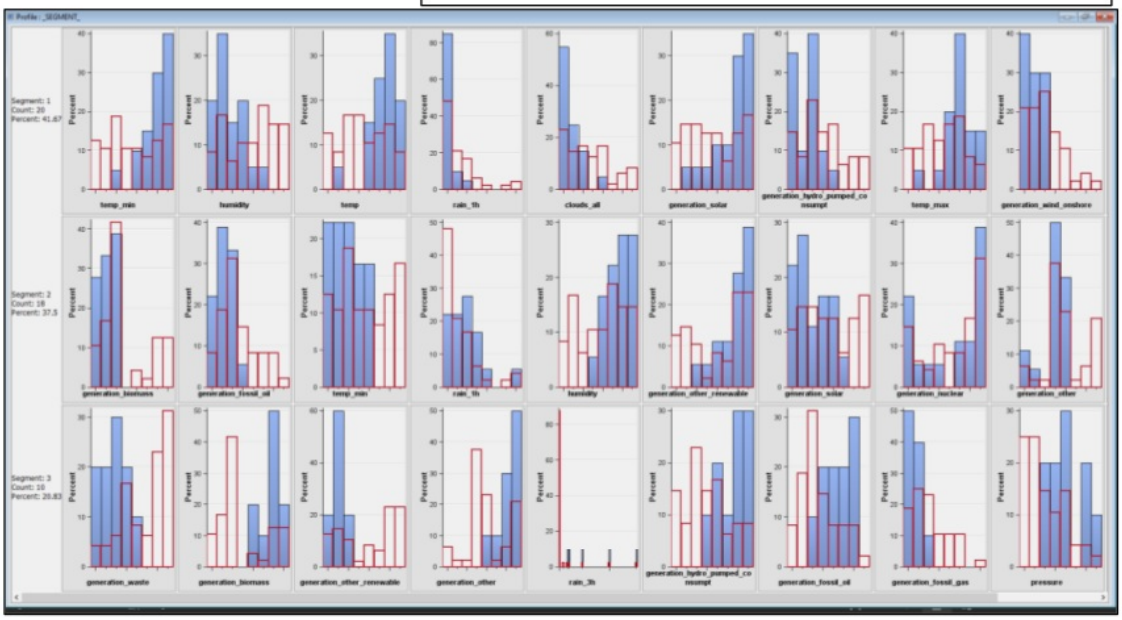


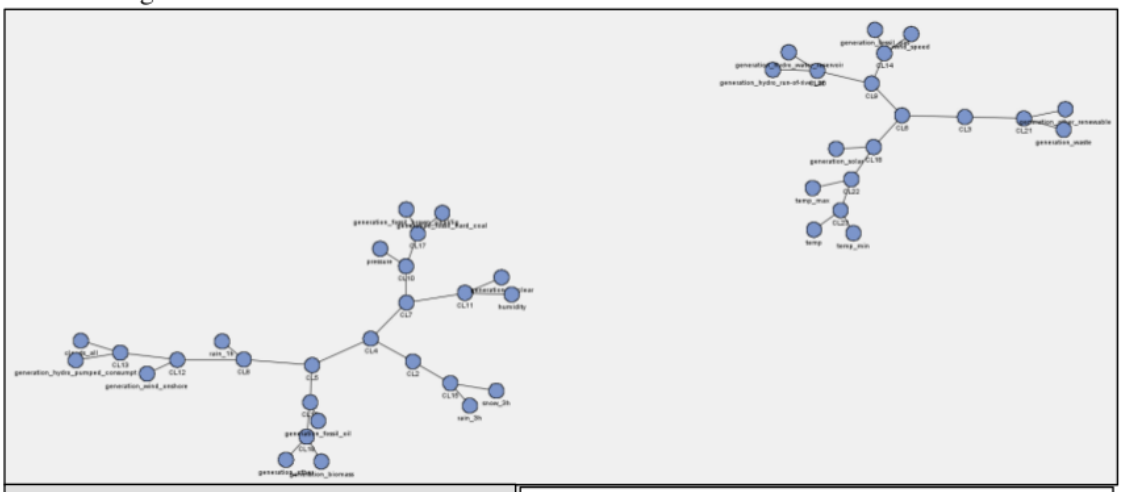
Figure 17: Cluster constellation, segments, variable importance and segment profiles for Seville data

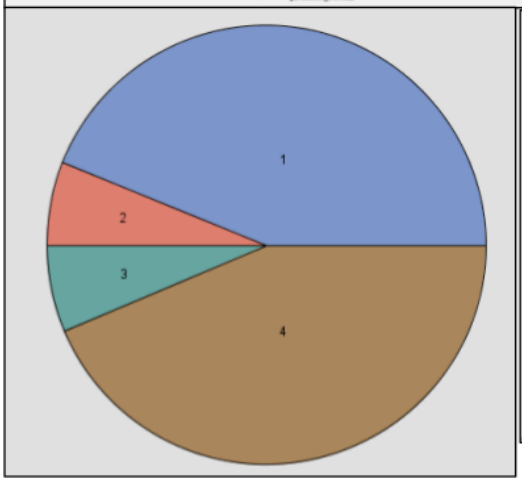
6. Critical Interpretation of Outcomes

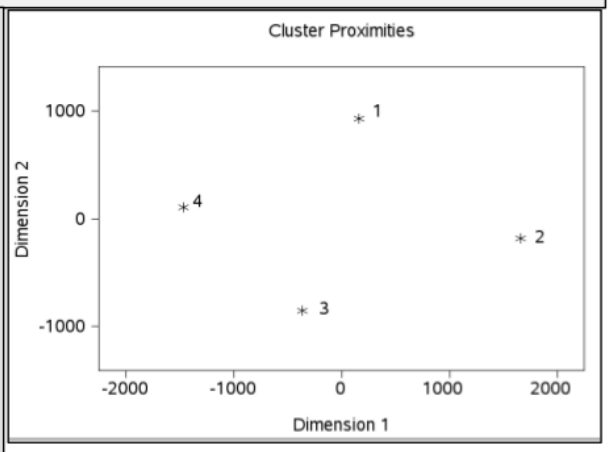
6.1. Interpretation of Similarity and Cluster Analysis

The first two figures in Figure 17 show the cluster constellation obtained from grouping up similar time series in the dataset. Here, one segment is clearly separated from the rest and seems to be characterized by temperature, *generation_solar*, *generation_fossil_gas* and others. This is the segment that is interesting to see. In addition, the variable importance and segment profile also rank high temperatures, low humidity, low rainfall, high *generation_solar*, low *clouds_all* and low *generation_wind_onshore* variables to be among the most important in the largest Segment 1. Segment 2, the second largest, is opposite to Segment 1, in which other energy sources such as biomass, fossil oil and nuclear are more representative, with low to moderate temperature, high humidity, moderate rainfall and low to moderate *generation_solar*. For Segment 3, the interpretation becomes ambiguous because there are only energy variables but no weather variables to profile it.

Bilbao Data – For the Bilbao dataset, monthly aggregation and four clusters were used, producing the following results:







Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
temp_max		0	2	1.00000
temp_min		1	0	0.89280
temp		0	1	0.87179
generation_solar		0	1	0.83929
wind_speed		0	1	0.82817
generation_hydro_pumped_consumpt		0	1	0.80548
generation_other_renewable		0	2	0.67844
rain_1h		1	0	0.50595
humidity		0	1	0.49530
snow_3h		0	1	0.49530
generation_fossil_oil		1	0	0.49023
rain_3h		0	1	0.48988
generation_biomass		0	1	0.46936
generation_hydro_run_of_river_an	generation_hydro_run-of-river_an	0	1	0.45857
generation_fossil_hard_coal		0	1	0.45857
generation_other		0	1	0.45857

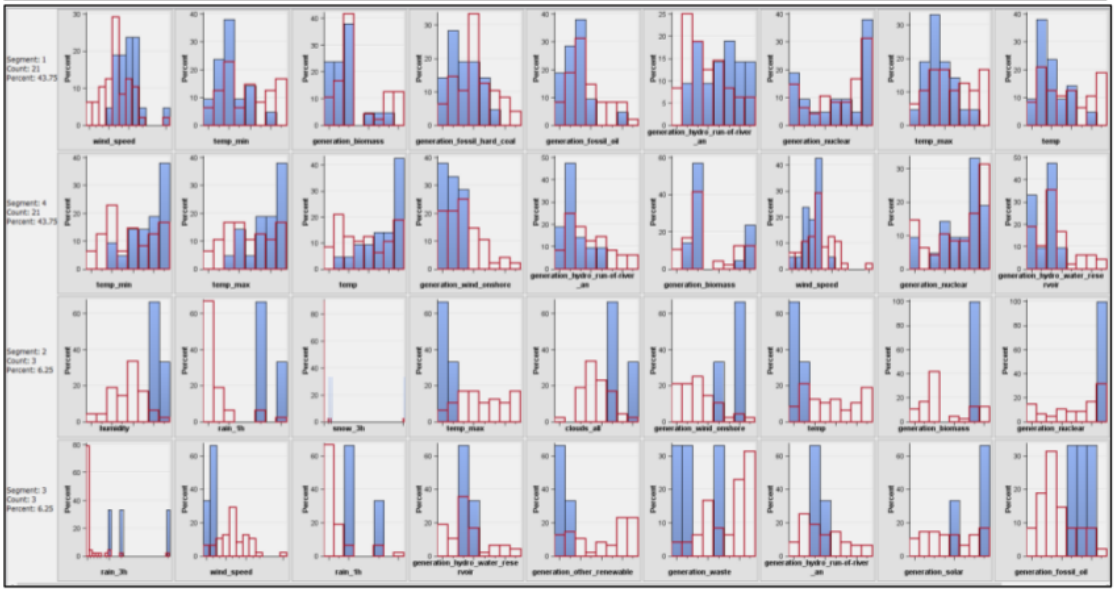
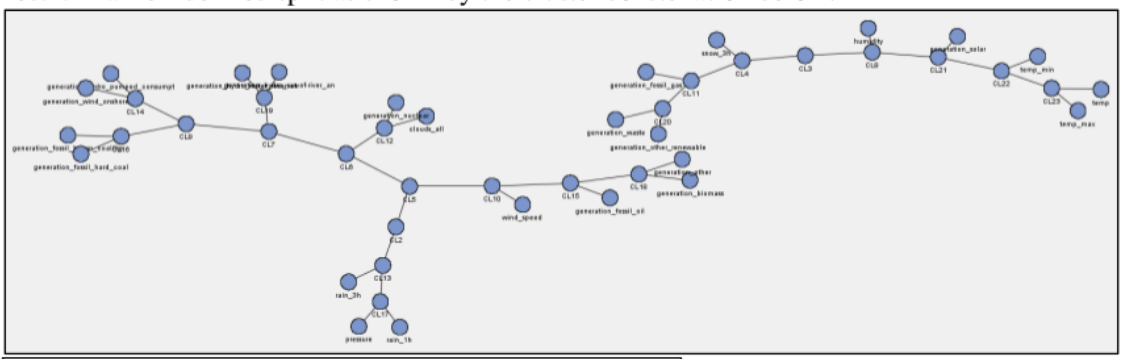


Figure 18: Cluster constellation, segments, variable importance and segment profiles for Bilbao data

The cluster constellation for Bilbao is segmented nearly equally into two parts, reflected by the two large clusters in the cluster diagram below it. In addition, there are also two small clusters, Segments 2 and 3, which are significantly far from the main clusters as well. Segment 1 is characterized by low-moderate temperature, low fossil fuel generation and moderate-high hydro-power generation. Segment 2 is characterized by high temperatures, low windspeed and wind-power generation but a higher instance of nuclear energy generation. Segment 3 is characterized by high humidity, precipitation and cloud cover, with very low temperatures but quite high instances of biomass and nuclear energy generation. This is contrary to the expectation that high

precipitation could drive a higher amount of hydro-power. Segment 4 is characterized by a high usage of fossil oil but other than that, the interpretation is ambiguous.

Barcelona Data – The Barcelona dataset, with monthly aggregation and five clusters, did not result in a well-defined split as shown by the cluster constellation below.



Variable Name	Label	Rules	Rules	Importance
generation_solar		1	0	1.00000
temp		0	1	0.98198
temp_max		0	1	0.98198
temp_min		0	1	0.96362
generation_other		0	1	0.93124
generation_other_renewable		0	1	0.93124
generation_biomass		1	0	0.93124
generation_hydro_pumped_consumpt		0	1	0.92582
generation_waste		0	1	0.91981
generation_fossil_oil		0	1	0.91981
generation_wind_onshore		0	1	0.90633
wind_speed		0	1	0.88465
generation_fossil_brown_coal_lig	generation_fossil_brown_coal/lig	1	0	0.67955
generation_fossil_hard_coal		0	1	0.67243
pressure		0	1	0.66524
generation_hydro_water_reservoir		0	1	0.66524
generation_fossil_gas		0	1	0.65062
generation_hydro_run_of_river_an	generation_hydro_run-of-river_an	0	1	0.65062

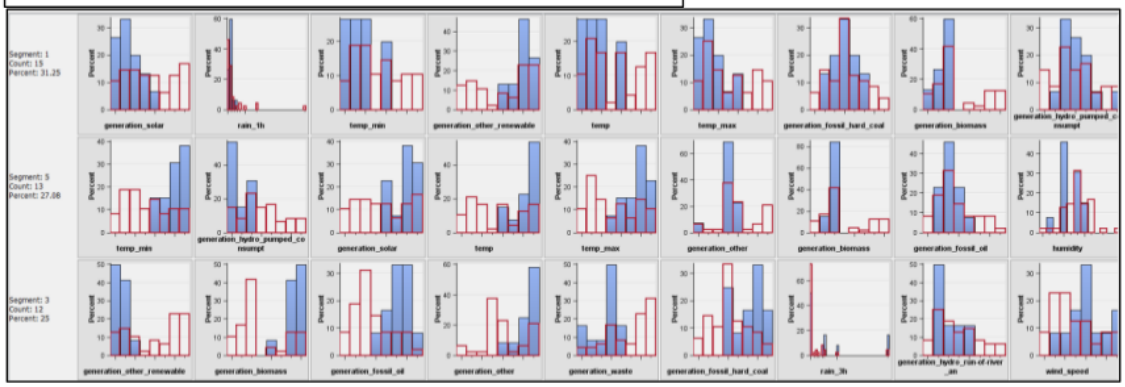


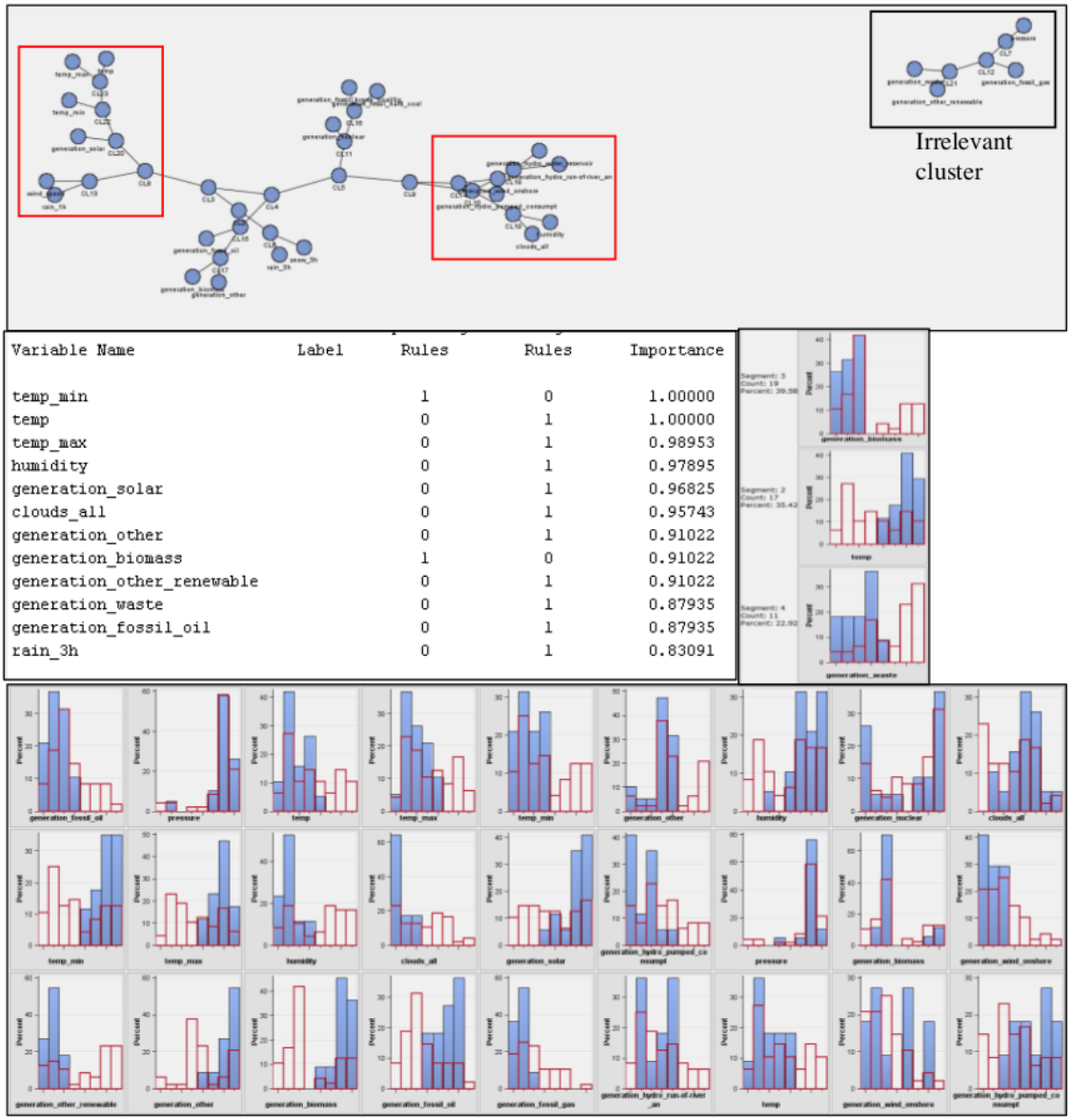
Figure 19: Cluster constellation, variable importance and segment profiles for Barcelona data

However, the variable importance table and segment profiles indicate several possible renewable energies such as solar energy, pumped-storage hydropower and wind-power as well as temperature and wind speed to be among the most important variables. Similar to the largest Seville segment, Segment 5 is characterized by high temperatures, higher proportion of solar energy generation, and low generation of energy from fossil oil and pumped-storage hydropower. In contrast,

Segment 1 is characterized by higher proportion of other renewables and pumped-storage hydropower, where temperatures are lower and there is lower production of solar energy.

The non-conclusive split of the segments could indicate that several possible energy generators are suitable in Barcelona, but their usefulness/gain would not be as substantial or long-lasting due to the smaller cluster sizes.

Madrid Data – The Madrid cluster constellation (with five clusters) shows two promising clusters on the left and an irrelevant cluster on the right (does not contain renewables/weather information).



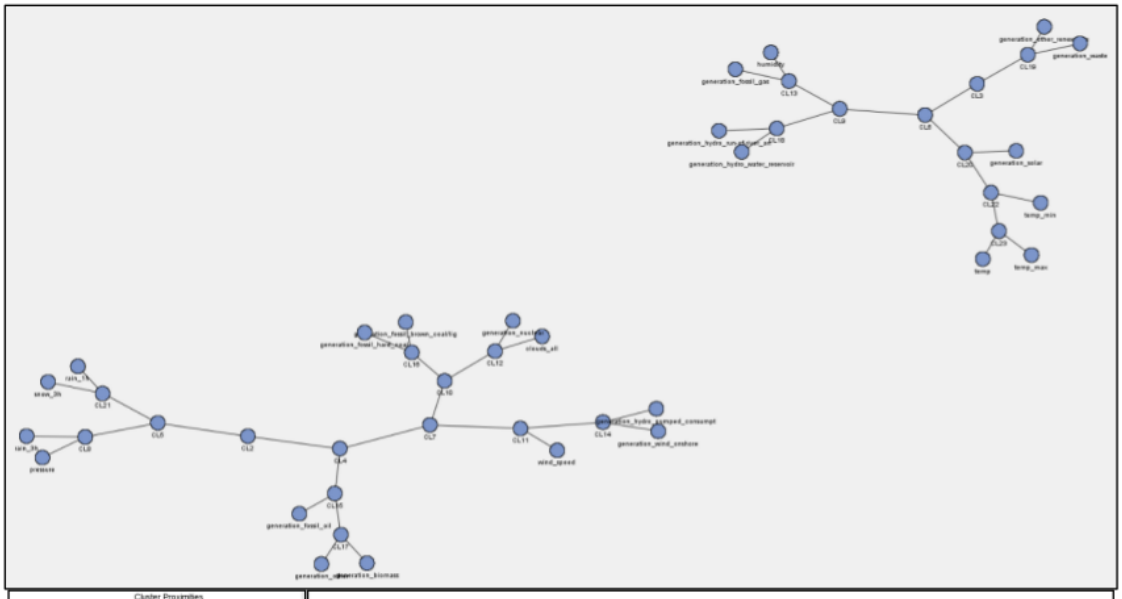
Variable Name	Label	Rules	Rules	Importance
temp_min		1	0	1.00000
temp		0	1	1.00000
temp_max		0	1	0.98953
humidity		0	1	0.97895
generation_solar		0	1	0.96825
clouds_all		0	1	0.95743
generation_other		0	1	0.91022
generation_biomass		1	0	0.91022
generation_other_renewable		0	1	0.91022
generation_waste		0	1	0.87935
generation_fossil_oil		0	1	0.87935
rain_3h		0	1	0.83091

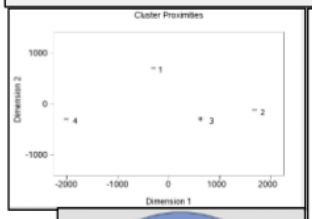
Figure 20: Cluster constellation, variable importance and segment profiles for Madrid data

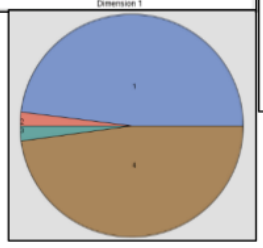
The leftmost highlighted cluster could be called the “temperature and solar energy” cluster, while the middle highlighted cluster could be called the “hydropower” cluster. Upon inspecting the segment profiles, the second largest segment, with high temperatures and pressure, low cloud cover and wind energy generation can be attributed to the “temperature and solar energy” cluster, while the third largest segment, with moderate-high production of wind and hydropower energies, can be attributed to the “hydropower” cluster.

Once again, this is similar to the Barcelona case in which several renewable energy and weather combinations are possible, and as a result none of them would yield a substantial amount of energy compared to regions with more sustained and definitive weather conditions.

Valencia Data – The Valencia cluster constellation was created with four clusters, showing a definite split into two major segments. The two main segments show a significant separation







Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
generation_hydro_run_of_river_an	generation_hydro_run-of-river_an	0	2	1.00000
generation_hydro_water_reservoir		0	2	0.99910
wind_speed		1	0	0.81883
generation_hydro_pumped_consumpt		0	1	0.74748
generation_wind_onshore		0	1	0.73808
temp		0	1	0.72856
generation_fossil_hard_coal		1	0	0.71014
generation_biomass		0	1	0.69896
generation_fossil_oil		0	1	0.66428
generation_fossil_brown_coal_lig	generation_fossil_brown_coal/lig	0	1	0.66428

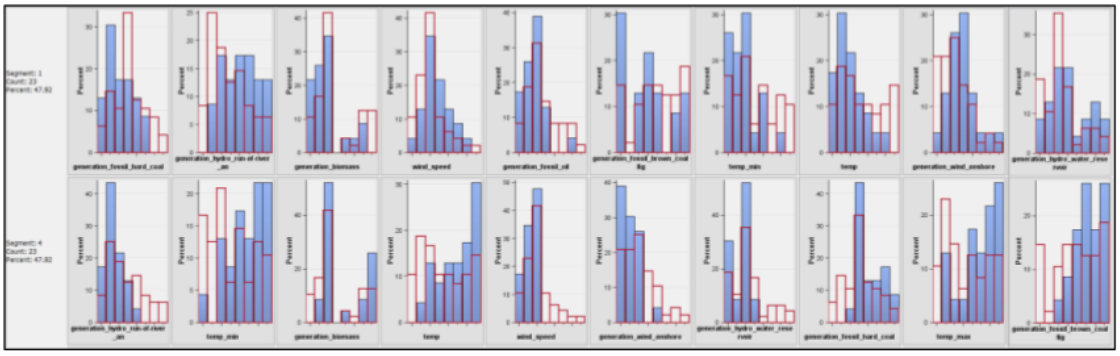


Figure 21: Cluster constellation, variable importance, segments and segment profiles for Valencia data

between them and are characterized by hydropower energy and wind energy. The most important variables are also seen to be hydropower energies, wind speed and wind energy generation. Both segments are of equal size, with Segment 1 being characterized by run-of-river hydropower and moderate-high wind speeds and wind energy generation, while Segment 4 is characterized by higher generation of fossil fuel energy when the wind speed, wind energy generation and hydropower generation are low. This could be interpreted that when renewable energy generation fails to deliver (perhaps due to adverse weather conditions), then fossil fuel usage is going to be higher, and vice versa. Nevertheless, for Valencia, both hydropower and wind power are possible implementations, with hydropower being slightly more favored.

The complete process flow in its entirety is shown in Figure 22 below:

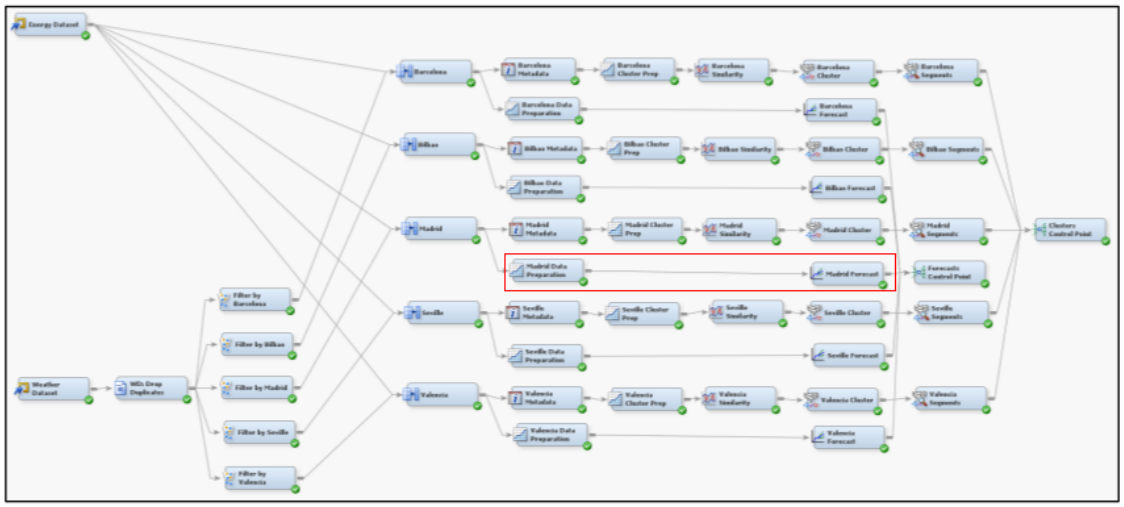
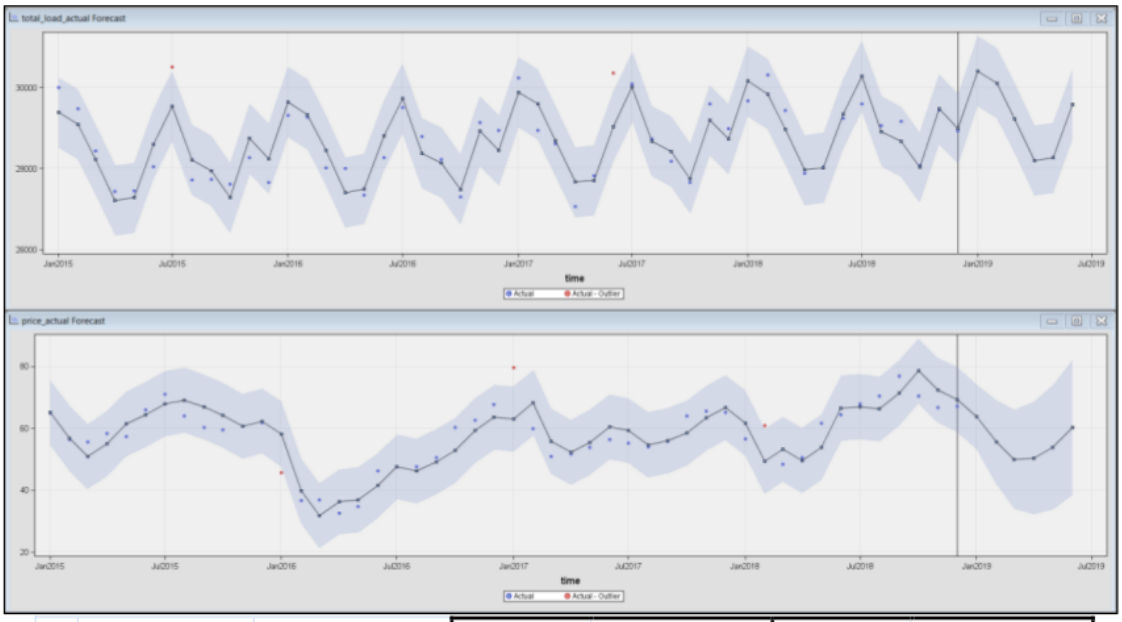


Figure 22: The complete process flow, in its entirety in SAS Enterprise Miner. The main time series forecast is highlighted in the red box (the Madrid forecast).

6.2. Interpretation of Price and Load Forecasts

The monthly average forecast diagrams for price and load, as well as the table of forecast experiments for different aggregations and their error analyses are shown again in Figure 23 below for ease of reference and interpretation:



	Aggregation	Forecast lead	Load R^2	Load *MAPE*	Price R^2	Price *MAPE*
1	Hour	24	96.7%	1.99%	96.2%	3.96%
2	Day	7	26.5%	6.19%	84.5%	6.90%
3	Week	6	61.4%	2.37%	81.2%	6.27%
4	Semi-month	6	79.2%	1.47%	80.8%	6.27%
5	Month	6	78.5%	1.17%	73.8%	6.98%
6	Quarter	4	90.3%	0.488%	47.5%	9.69%

Figure 23: Forecast diagrams (monthly aggregation only) and error analyses from Section 5.1

It can be seen by looking at the *total_load_actual* (total energy demand by consumers) diagram, that the seasonality of the time series is 12 months and there is an increasing trend over time. The forecast for the next six months was able to accurately capture the shape of the periodic pattern, starting from the peak in January to the flatline in April-May and the following rise in June. If extrapolated further, the load pattern from July to December could also be captured. There are clear peaks in January and July, where weather is the most extreme and clear troughs in April-May and October, where the weather is milder. This confirms the hypothesis that the seasonality in consumer demand for energy is influenced by the seasonality in weather. Consumers demand more energy for heating appliances in the winter and for cooling appliances in the summer. Such appliances are mostly unnecessary during mild seasons such as spring and autumn.

In stark contrast to the predictability of energy load, the electricity price diagram has neither the seasonality nor the pattern portrayed by energy load. This means that there are other factors that drive electricity price to increase or decrease and cannot be solely attributed to weather or consumer demand. If this were not the case, then price would simply rise during high demand and fall during low demand, with a well-defined pattern similar to energy load. This can still be seen in the diagram when there are sometimes small rises and falls, but it doesn't explain the sustained increase/decrease in price over longer periods of time. The monthly forecast for price is also the second most uncertain, with a MAPE of almost 7%, compared to the monthly forecast for energy load, which has the second lowest MAPE at 1.17%.

Interestingly, the quarter-aggregated forecast for load is the most accurate, at only 0.5% MAPE and even has a 90% R^2 value. Perhaps the four quarters in each year is a reflection of the four seasons in weather, making it easy to accurately forecast energy load three months ahead of time. On the other hand, daily forecasts for load produced unexpectedly bad results, with quite a very low R^2 value. The reason for this can only be speculated that daily consumer demand for energy is unpredictable. Some days, demand could be low/high for reasons that could not be accounted for in this time series forecast. For finer intervals of time, hourly aggregated forecasts were expected to be the most accurate, and it appears to be so for both load and price. There is no clear seasonality or trend for price, so as the aggregation increases from daily to quarterly, the accuracy and fit for the price forecast decreases. For energy load, which is dependent on longer periods of time, the forecasts gave poor results for daily and weekly aggregation but improved for semi-monthly, monthly and quarterly aggregation.

Hence, it is recommended to forecast 24 hours ahead or several days into the future for best results. For energy load, it is recommended to forecast 24 hours ahead, semi-monthly, monthly or quarterly forecasts for best results. The implication for energy companies is that they may not be able to foresee in advance, sudden increases in daily or weekly load, and hence may not be able provide the required and timely energy supply for consumers. This can be one explanation for the lack of a clear pattern and unpredictability in price. Further discussion is done in Section 7: Conclusion and Discussion.

7. Discussion and Conclusion

Although the electric power industry is one of the biggest contributors to greenhouse gas emissions, it is also an industry that generates large amounts of data. The aim of REE, the energy service company in Spain, is to mine this data in order to discover ways to reduce wastage and implement renewable energy sources. More specifically, the aim of REE is to use energy and weather data to identify the optimal location out of five cities in Spain, in which to place a renewable energy generator. Along the way to achieving this aim, the objective of forecasting energy demand and electricity price was also set, so that the company can be more prepared to provide the required energy supply with minimum overproduction or underproduction.

All of the objectives and aims were achieved. The first objective of visualizing and describing trends in energy and weather was done by analyzing the two datasets on Tableau. The second objective of forecasting consumer energy demand (total load) and electricity price was done on SAS Enterprise Miner using exponential smoothing forecast models. The third and final objective of prescribing the optimal renewable energy generator locations was done by exploring energy-weather variable clusters through a time series similarity and cluster analysis of each city.

The forecast for energy load was found to be predictable, having both seasonality and trend, when aggregated over longer horizons (0.5 to 3 months) due to the seasonality of weather. On the other hand, the forecast for price was less predictable as the forecast horizon increases, with the best forecast being 24 hours ahead. The energy-weather time series clusters for Barcelona and Madrid were found to be inconclusive. The largest Seville cluster was found to favor conditions for solar energy generation, which includes low humidity, warm temperatures and low cloud cover. The Bilbao clusters were characterized by humidity, wind, rain, and cloud cover, wind-power and hydro-power generation. In addition, water reservoir and run-of-river hydropower generation was found to be clustered with moderate amounts of rainfall, although the cluster is quite small. A correlation between wind speed and wind-power generation was also found. Lastly, the Valencia clusters provided good evidence between wind speed and wind-power generation, as well as demonstrating the inverse relationship between non-renewable energy generation and hydropower/wind energy. That is, when renewables provide less energy due to poor weather conditions, fossil fuel energy generation is increased.

The reliance of renewables on the weather could perhaps be a reason for the unpredictability in price. If an energy company only uses renewable sources, the energy supply is overly reliant on optimal weather conditions. Even when a mix of both renewable and non-renewable sources are used, responding to sharp fluctuations in short-term energy demand is difficult because of the inherent difficulty in storing energy. However, these fluctuations smoothen out over the long run, as seen from the monthly average load forecasts. In short, the mismatch between the predictability of energy demand with the unpredictability of energy supply causes the unpredictability in electric price. The final recommendation would be to construct a solar farm in a location that approximates

Seville's weather conditions. Solar farms are much cheaper than hydroelectric power stations, and more predictable in terms of seasonality in weather. Unlike hydropower, solar farms can be used for a small-scale, more targeted approach rather than for city-wide consumption.

Several limitations are evident in this study. Firstly, the energy dataset contains data for the entire country of Spain while the weather dataset is only for the five largest cities in Spain. While this has an advantage of providing the means to link together energy and weather conditions, it also brings in unrepresentative city-specific clusters. For example, small clusters could still be found where high temperature and high solar energy generation are linked simply because that is the case during summertime. Secondly, more preprocessing and variable importance tests could have been done before carrying out similarity and cluster analysis. Finally, the cluster analyses were not as objective and clear-cut as it was originally intended to be, probably because choosing the number of clusters and interpreting them is more of an art than a science.

In conclusion, this project only scratched the surface of the enormous amount of analysis that could be done. In a future study, forecasts for each of the weather metrics and energy generation can be done and compared with company forecasts. Regression forecast models for load and price could be investigated by including some of the most important independent variables. In addition, if interpretation was not an issue, the option of transposing time series could be explored before similarity analysis and the output can be fed into the cluster analysis node, where more objective clusters can be created.

References

ENTSO-E, 2020. *ENTSO-E Transparency Platform*. [Online] Available at: https://transparency.entsoe.eu/dashboard/show [Accessed 16 July 2020].

Heggie, J., 2020. *Spain: taking sustainable energy to the next level*. [Online] Available at: https://www.nationalgeographic.com/science/2020/02/partner-content-setting-standard-for-sustainability/ [Accessed 17 July 2020].

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2 [Accessed 17 July 2020].

IPCC, 2014: *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jhana, N., 2019. *Hourly energy demand generation and weather | Kaggle*. [Online] Available at: https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather [Accessed 15 July 2020].

OpenWeather Ltd., 2020. *Weather API - OpenWeatherMap*. [Online] Available at: https://openweathermap.org/api [Accessed 17 July 2020].

REE, 2020. *Markets and prices | ESIOS electricity · data · transparency*. [Online] Available at: https://www.esios.ree.es/en/market-and-prices [Accessed 16 July 2020].

Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A.S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E.D., Mukkavilli, S.K., Kording, K.P., Gomes, C., Ng, A.Y., Hassabis, D., Platt, J.C., Creutzig, F., Chayes, J. & Bengio, Y. (2019). Tackling Climate Change with Machine Learning. *CoRR*. [Online]. abs/1906.05433. Available from: http://arxiv.org/abs/1906.05433

Solucion Asesores XXI, 2016. *WHY SPAIN - Solucion Asesores XXI*. [Online] Available at: https://solucionasesoressl.com/en/why-spain [Accessed 16 July 2020].

ABAV Assignment C

ORIGINALITY REPORT

2%	%	%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1%
2	madoc.bib.uni-mannheim.de Internet Source	<1%
3	Submitted to Swinburne University of Technology Student Paper	<1%
4	Submitted to University of Technology, Sydney Student Paper	<1%
5	essay.utwente.nl Internet Source	<1%
6	www.kaggle.com Internet Source	<1%
7	www.unofficialgoogledatascience.com Internet Source	<1%
8	telerehab.pitt.edu Internet Source	<1%

9	www.moef.gov.in Internet Source	<1%
10	www.wri.org Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

ABAV Assignment C

GRADEMARK REPORT

FINAL GRADE

43 /50

GENERAL COMMENTS

Instructor

Data prep and EDA - Good initiative to ensure quality of data. Cleaning, filtering, merging of data observed.

2 models - Time series and clustering are explored and well documented. The models ae validated and optimized.

Overall process flows reveals several replicas of the same model for different locations.

Interpretation of outcomes was appropriate and suitable conclusions have been drawn.

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

Comment 1 | Data Prep

Good observation. and suitable prep

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

RUBRIC: ABAV_PARTC_JUNE 2020 4.30 / 5

DATA PREP (20%) 4 / 5

Data Preparation

DISTINCTION
(5)

ABOVE AVERAGE
(4)

AVERAGE
(3)

BELOW AVERAGE
(2)

FAIL
(1)

MODEL (30%) 5 / 5

Model Construction, Optimization and Validation

DISTINCTION
(5)

ABOVE AVERAGE
(4)

AVERAGE
(3)

BELOW AVERAGE
(2)

FAIL
(1)

OUTCOME (30%) 4 / 5

Critical Interpretation of Outcomes

DISTINCTION
(5)

ABOVE AVERAGE
(4)

AVERAGE
(3)

BELOW AVERAGE
(2)

FAIL
(1)

CONCLUSION (20%) 4 / 5

Discussion and Conclusion

DISTINCTION
(5)

ABOVE AVERAGE
(4)

AVERAGE
(3)

BELOW AVERAGE
(2)

FAIL
(1)