COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

Spanish Electricity Market Analytics with Python

Prof. Andres Ramos
https://www.iit.comillas.edu/aramos/
Andres.Ramos@comillas.edu

February 2024

comillas.edu

Content

1. Iberian Electricity Market
2. Data Description
3. Data Analysis. ANOVA
4. Reducing dimensions
5. Representative patterns
6. Price estimation

1

Iberian Electricity Market

Spanish Electricity Market

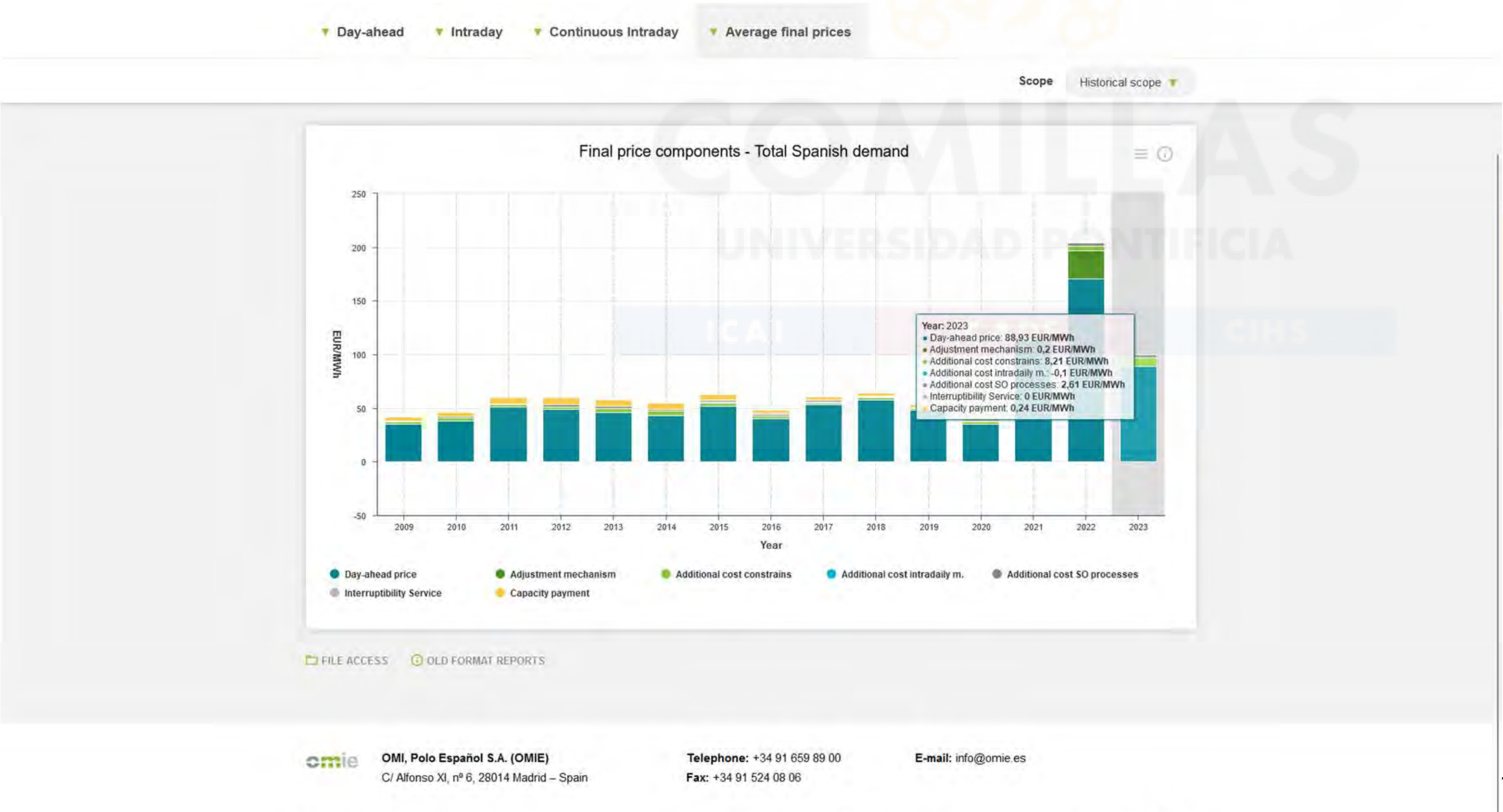
omie

- OMIE (www.omie.es/en/) manages the wholesale electricity market (referred to as cash or "spot") on the Iberian Peninsula. Like any other, the electricity market caters to trading electricity between agents (producers, consumers, retailers, etc.) at a price that is known, transparent, and accessible. At OMIE, we guarantee that the market is operated in a transparent and non-discriminatory manner. In January 1998, we began our operations for the Spanish market, and in July 2007 we extended them to cover the whole Iberian Market.

red eléctrica

- According to www.esios.ree.es/en, the role of REE as the System Operator consists of maintaining the balance generation-consumption and, for this purpose, it produces the electricity demand forecasts, oversees the operation of the generation facilities and manages the transmission facilities in real-time, constantly ensuring that scheduled generation in power stations matches consumer demand.

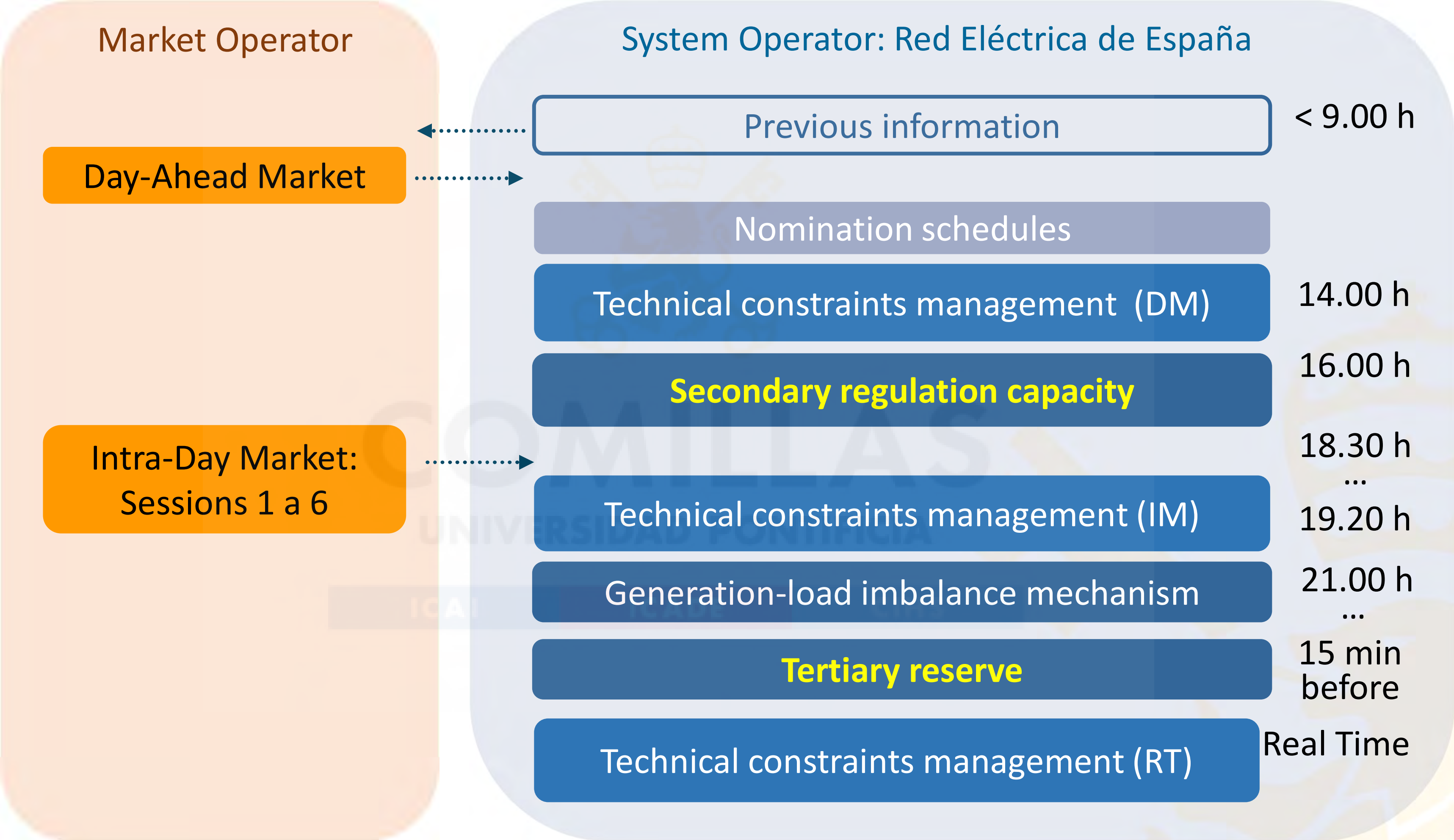
Market Products



https://www.omie.es/en/market-results/interannual/average-final-prices/components-spanish-demand?scope=interannual

- Market Operator (MO)
  - Daily Market
  - Intraday Market
- System Operator (SO)
  - Interruptibility Service
  - Capacity Payment
  - Adjustment Services
    - Technical constraints
    - Secondary reserve
    - Tertiary reserve
    - Upward reserve power
    - Imbalances

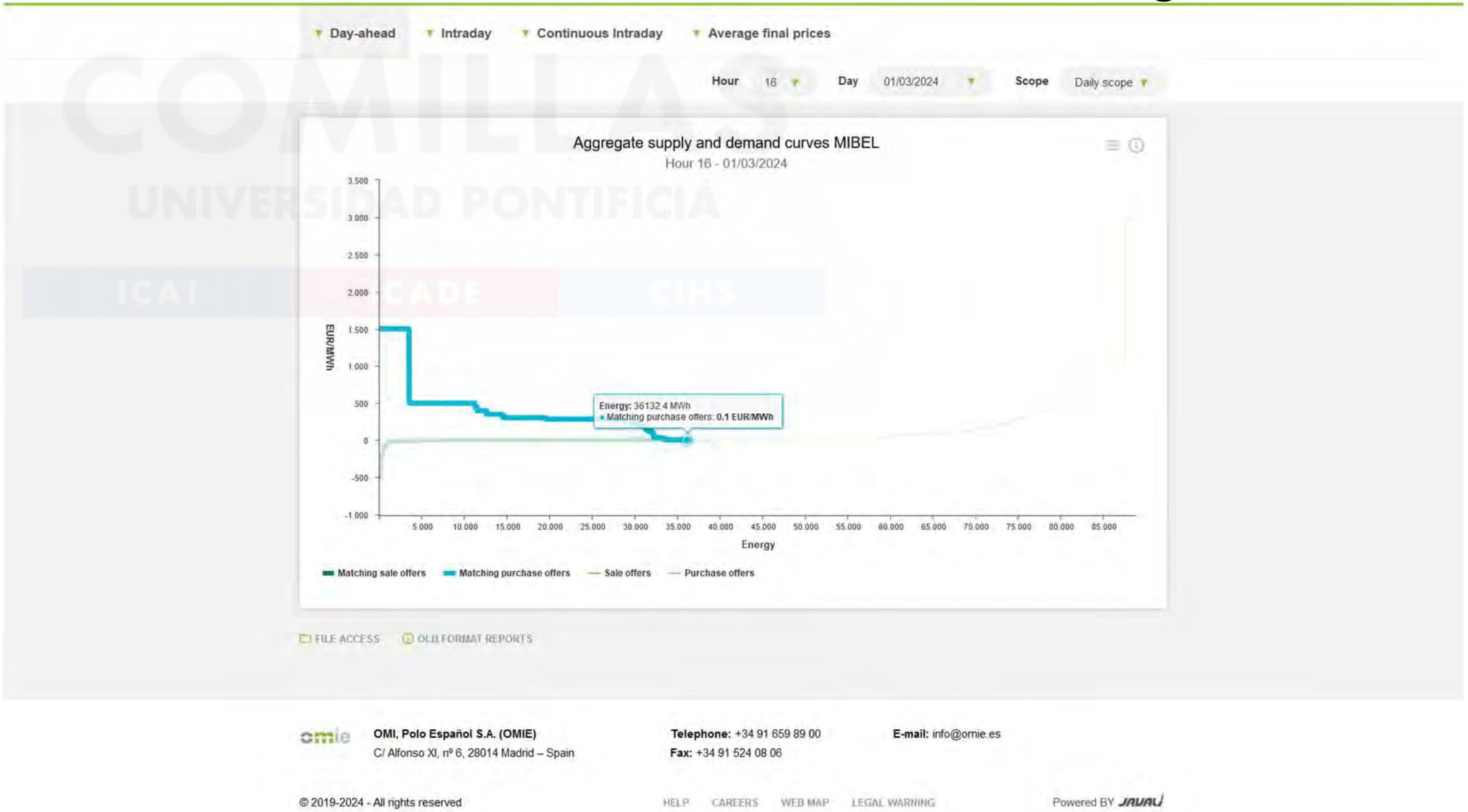
Short-term operation scheduling



Source: M. de la Torre, J. Paradinas *Integration of renewable generation. The case of Spain*

Spanish Daily Electricity Market

- In the day-ahead market, buying and selling agents submit their purchase and sale bids for the twenty-four hours of the following day. The price and volume of energy over a specific hour are determined by the point at which the supply and demand curves meet, according to the marginal pricing model adopted by the EU, based on the algorithm approved for all European markets (EUPHEMIA). The scheduled generation of each unit is also obtained as a result of this algorithm.



YEAR
MONTH
DAY
HOUR
DAYWEEK

Hydro_UGH
Hydro_non_UGH
Turbine_pumping
Nuclear
Soft_coal/Anthracite
Sub-bituminous_coal
Combined_cycle_GT
Fuel
Natural_Gas
Onshore_wind
Solar_PV
Solar_thermal
Geothermal_and_Ocean
Natural_Gas_Cogeneration
Fossil_Oil
Mining_subproducts
Residual_energy
Biomass
Biogas
Household_and_similar_wastes_
Sundry_waste
Pump_consumption
Balearic_HVDC_Link
Generic
Fuel-Gas
Cogeneration
Other_renewables

Portugal_balance
France_balance
Morocco_balance
Andorra_balance

Total_scheduled_demand
Total_Program_Generation

Average_price_final

Dataset

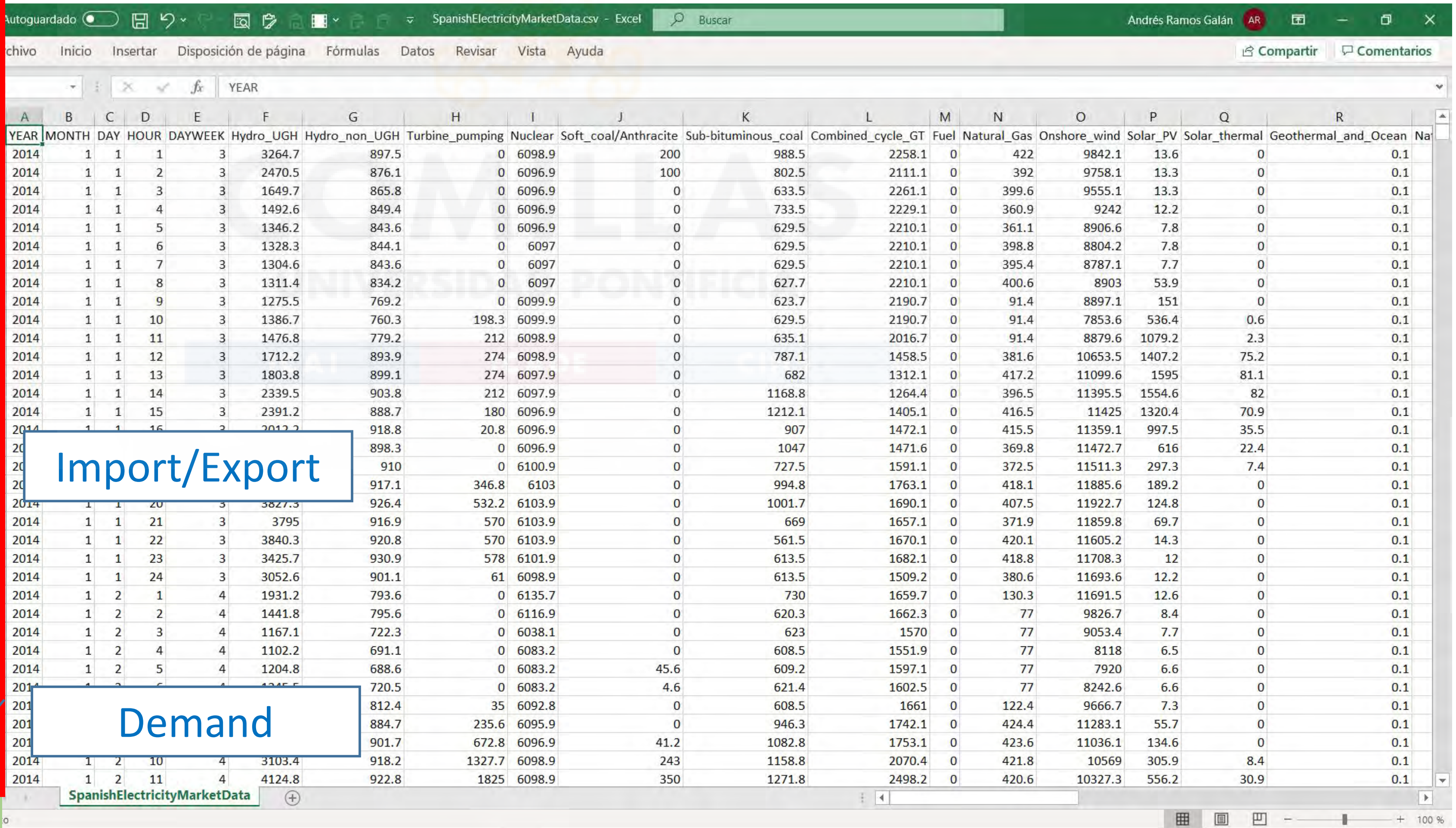
Hourly data

- Data from 2014/01/01 to 2019/12/31 (6 years)

Technologies

Import/Export

Demand



YEAR	MONTH	DAY	HOUR	DAYWEEK	Hydro_UGH	Hydro_non_UGH	Turbine_pumping	Nuclear	Soft_coal/Anthracite	Sub-bituminous_coal	Combined_cycle_GT	Fuel	Natural_Gas	Onshore_wind	Solar_PV	Solar_thermal	Geothermal_and_Ocean
2014	1	1	1	3	3264.7	897.5	0	6098.9	200	988.5	2258.1	0	422	9842.1	13.6	0	0.1
2014	1	1	2	3	2470.5	876.1	0	6096.9	100	802.5	2111.1	0	392	9758.1	13.3	0	0.1
2014	1	1	3	3	1649.7	865.8	0	6096.9	0	633.5	2261.1	0	399.6	9555.1	13.3	0	0.1
2014	1	1	4	3	1492.6	849.4	0	6096.9	0	733.5	2229.1	0	360.9	9242	12.2	0	0.1
2014	1	1	5	3	1346.2	843.6	0	6096.9	0	629.5	2210.1	0	361.1	8906.6	7.8	0	0.1
2014	1	1	6	3	1328.3	844.1	0	6097	0	629.5	2210.1	0	398.8	8804.2	7.8	0	0.1
2014	1	1	7	3	1304.6	843.6	0	6097	0	629.5	2210.1	0	395.4	8787.1	7.7	0	0.1
2014	1	1	8	3	1311.4	834.2	0	6097	0	627.7	2210.1	0	400.6	8903	53.9	0	0.1
2014	1	1	9	3	1275.5	769.2	0	6099.9	0	623.7	2190.7	0	91.4	8897.1	151	0	0.1
2014	1	1	10	3	1386.7	760.3	198.3	6099.9	0	629.5	2190.7	0	91.4	7853.6	536.4	0.6	0.1
2014	1	1	11	3	1476.8	779.2	212	6098.9	0	635.1	2016.7	0	91.4	8879.6	1079.2	2.3	0.1
2014	1	1	12	3	1712.2	893.9	274	6098.9	0	787.1	1458.5	0	381.6	10653.5	1407.2	75.2	0.1
2014	1	1	13	3	1803.8	899.1	274	6097.9	0	682	1312.1	0	417.2	11099.6	1595	81.1	0.1
2014	1	1	14	3	2339.5	903.8	212	6097.9	0	1168.8	1264.4	0	396.5	11395.5	1554.6	82	0.1
2014	1	1	15	3	2391.2	888.7	180	6096.9	0	1212.1	1405.1	0	416.5	11425	1320.4	70.9	0.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	918.8	20.8	6096.9	0	907	1472.1	0	415.5	11359.1	997.5	35.5	0.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	898.3	0	6096.9	0	1047	1471.6	0	369.8	11472.7	616	22.4	0.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	910	0	6100.9	0	727.5	1591.1	0	372.5	11511.3	297.3	7.4	0.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	917.1	346.8	6103	0	994.8	1763.1	0	418.1	11885.6	189.2	0	0.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	926.4	532.2	6103.9	0	1001.7	1690.1	0	407.5	11922.7	124.8	0	0.1
2014	1	1	21	3	3795	916.9	570	6103.9	0	669	1657.1	0	371.9	11859.8	69.7	0	0.1
2014	1	1	22	3	3840.3	920.8	570	6103.9	0	561.5	1670.1	0	420.1	11605.2	14.3	0	0.1
2014	1	1	23	3	3425.7	930.9	578	6101.9	0	613.5	1682.1	0	418.8	11708.3	12	0	0.1
2014	1	1	24	3	3052.6	901.1	61	6098.9	0	613.5	1509.2	0	380.6	11693.6	12.2	0	0.1
2014	1	2	1	4	1931.2	793.6	0	6135.7	0	730	1659.7	0	130.3	11691.5	12.6	0	0.1
2014	1	2	2	4	1441.8	795.6	0	6116.9	0	620.3	1662.3	0	77	9826.7	8.4	0	0.1
2014	1	2	3	4	1167.1	722.3	0	6038.1	0	623	1570	0	77	9053.4	7.7	0	0.1
2014	1	2	4	4	1102.2	691.1	0	6083.2	0	608.5	1551.9	0	77	8118	6.5	0	0.1
2014	1	2	5	4	1204.8	688.6	0	6083.2	45.6	609.2	1597.1	0	77	7920	6.6	0	0.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	720.5	0	6083.2	4.6	621.4	1602.5	0	77	8242.6	6.6	0	0.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	812.4	35	6092.8	0	608.5	1661	0	122.4	9666.7	7.3	0	0.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	884.7	235.6	6095.9	0	946.3	1742.1	0	424.4	11283.1	55.7	0	0.1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	901.7	672.8	6096.9	41.2	1082.8	1753.1	0	423.6	11036.1	134.6	0	0.1
2014	1	2	10	4	3103.4	918.2	1327.7	6098.9	243	1158.8	2070.4	0	421.8	10569	305.9	8.4	0.1
2014	1	2	11	4	4124.8	922.8	1825	6098.9	350	1271.8	2498.2	0	420.6	10327.3	556.2	30.9	0.1

Business Analytics Spectrum

- **Descriptive**: statistics (data analysis, analysis of variance, correlation)
- **Predictive**: simulation, regression, forecasting
- **Prescriptive**: optimization, heuristics, decision analysis

Competitive advantage

Stochastic Optimization	How can we achieve the best outcome including the effects of variability?	*PRESCRIPTIVE*
Optimization	How can we achieve the best outcome?	
Predictive modeling	What will happen next if?	*PREDICTIVE*
Forecasting	What if these trends continue?	
Simulation	What could happen…?	
Alerts	What actions are needed?	
Query/drill down	What exactly is the problem?	*DESCRIPTIVE*
Ad hoc reporting	How many, how often, where?	
Standard reporting	What happened?	

Source: A. Fleischer et al. *ILOG Optimization for Collateral Management*

Unveiling the essential machine learning algorithms for data science in 2024

https://www.analyticsinsight.net/essential-machine-learning-algorithms-for-data-science-in-2024/

1. Linear regression: A method for supervised learning that forecasts a continuous variable as an output, given one or more input parameters. In regression issues like sales forecasting or home price estimation, it is one of the most straightforward and popular techniques. Assuming a linear connection between the input and output variables, it determines the line that fits the data the best and reduces the error between the predicted and actual values.

2. Logistic regression: A method for supervised learning that uses one or more input variables to predict a binary output variable. It is among the most widely used algorithms for categorization issues, including the identification of spam emails and illness diagnosis. It models the likelihood that an input belongs to a particular class using a logistic function, and then it applies a threshold to determine the outcome.

3. Decision tree: Supervised learning, using criteria to build decision-like tree structures, is a key Machine Learning method in 2024. Handling numerical and categorical data applies to regression and classification problems. Its simplicity mirrors human logic, making it easy to understand. However, overfitting can limit its generalization ability by capturing excessive noise and complexity.

4. Random forest: Random Forest, a supervised learning algorithm, combines multiple decision trees for a robust model. As an ensemble method, it merges predictions from various base models, enhancing performance. It introduces randomness into decision trees by using different data subsets and features, then averages or votes their predictions. This reduces overfitting and increases model accuracy and stability.

5. K-means clustering: An unsupervised learning algorithm clusters data points based on similarity. Popular for customer segmentation and image compression, it initializes cluster centers randomly, assigns data points to the nearest cluster, and updates centers until convergence. However, it's sensitive to initial cluster centers, cluster numbers, and data outliers.

6. Support vector machine (SVM): A supervised learning system called Support Vector Machine efficiently divides data points into several groups. It works well for classification issues, particularly when dealing with non-linear, high-dimensional data. It converts data into a higher-dimensional space for simpler linear separation using the kernel method. It manages regression, binary, and multi-class issues.

7. Apriori: In transactional databases, a technique known as unsupervised learning finds common itemsets and association rules. It studies client purchase habits and is often used for market basket analysis. From a bottom-up approach, it generates candidates by applying minimal support and confidence levels to reject out-of-date itemsets and rules.

8. Artificial neural network (ANN): A supervised learning technique called a neural network uses linked neurons to simulate the structure of the brain. It is an intricate Data Science system that can learn from any type of data and accomplish tasks like speech synthesis, picture recognition, and natural language processing. Based on input from the output, it modifies the weights and biases of neuron connections to adapt.

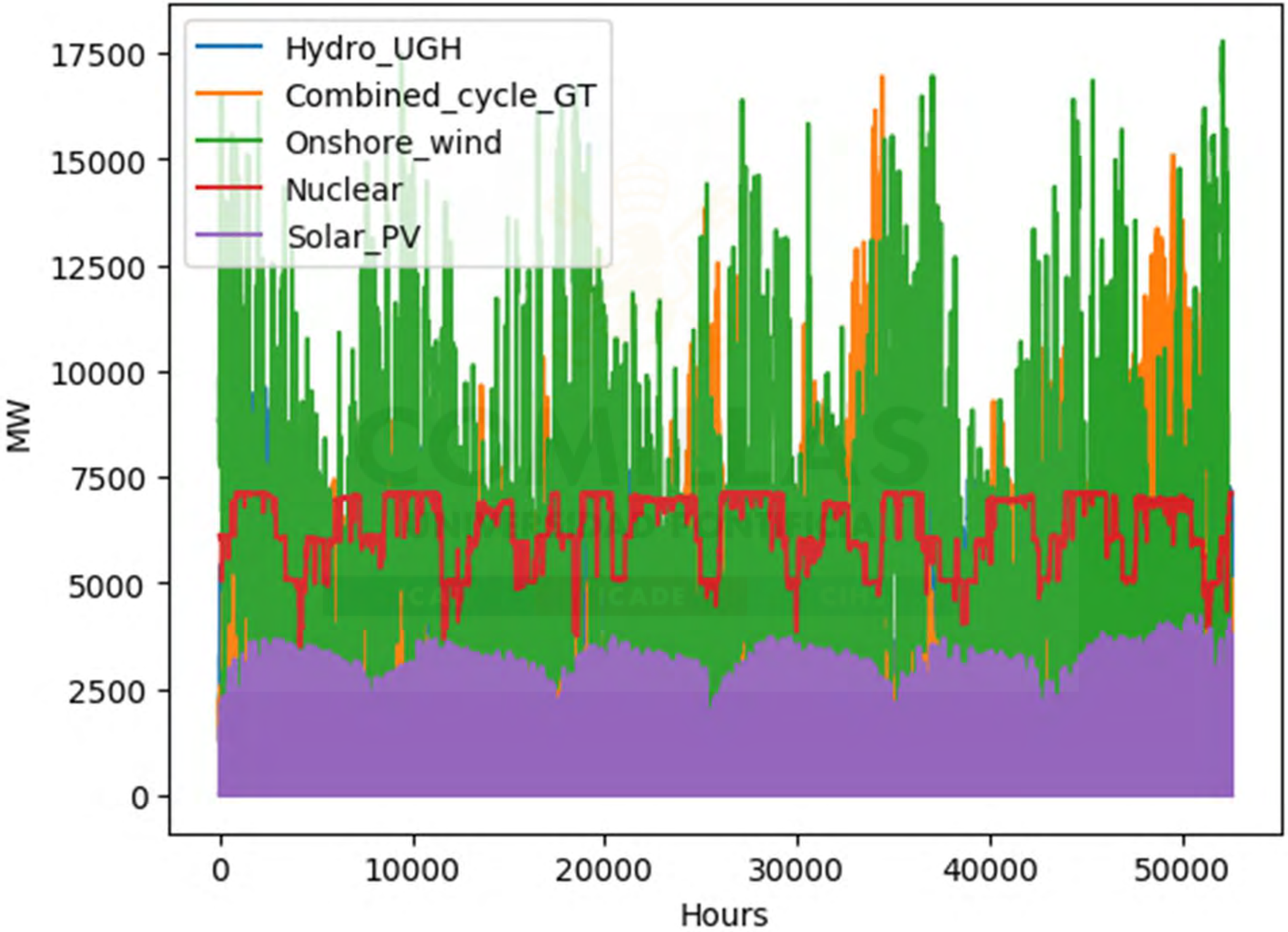
9. K-nearest neighbors (KNN): The supervised learning method K-nearest neighbors make predictions about outputs by using the k-closest neighbors found in the training set. It examines the similarity or distance between data points, and then takes the average of the results or a majority vote, making it perfect for regression and classification. It is susceptible to the choice of k and distance metric, though, and can be computationally costly.

10. Naïve Bayes: A supervised learning technique called Naive Bayes forecasts results using the prior output probability and the conditional probability of features. It is predicated on the feature independence of Bayes' theorem. For classification tasks, particularly text analysis, it is quick and easy to use. But if the data deviates from the independence assumption or the prior probability isn't representative, it can be wrong.

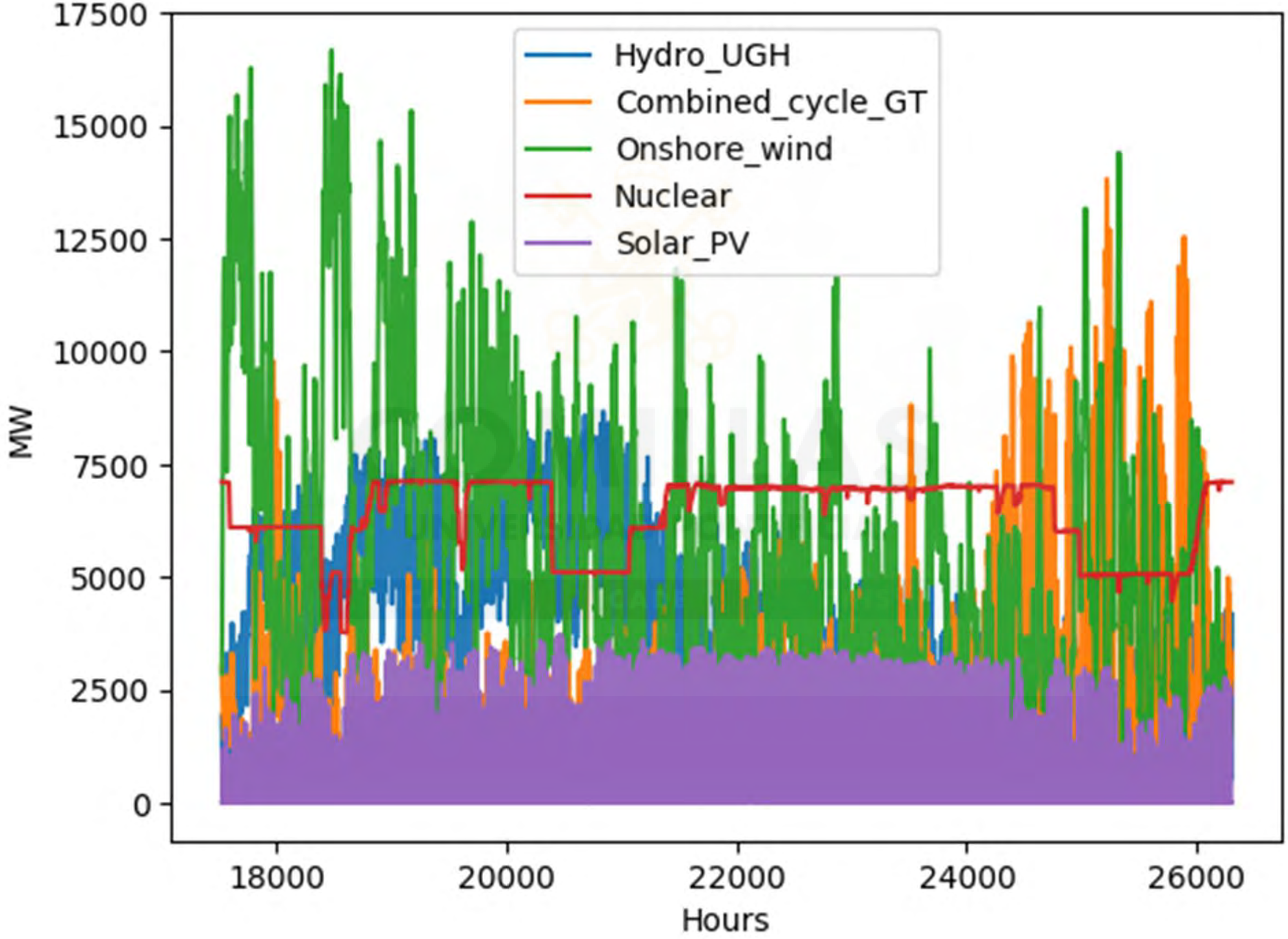
2

Data description

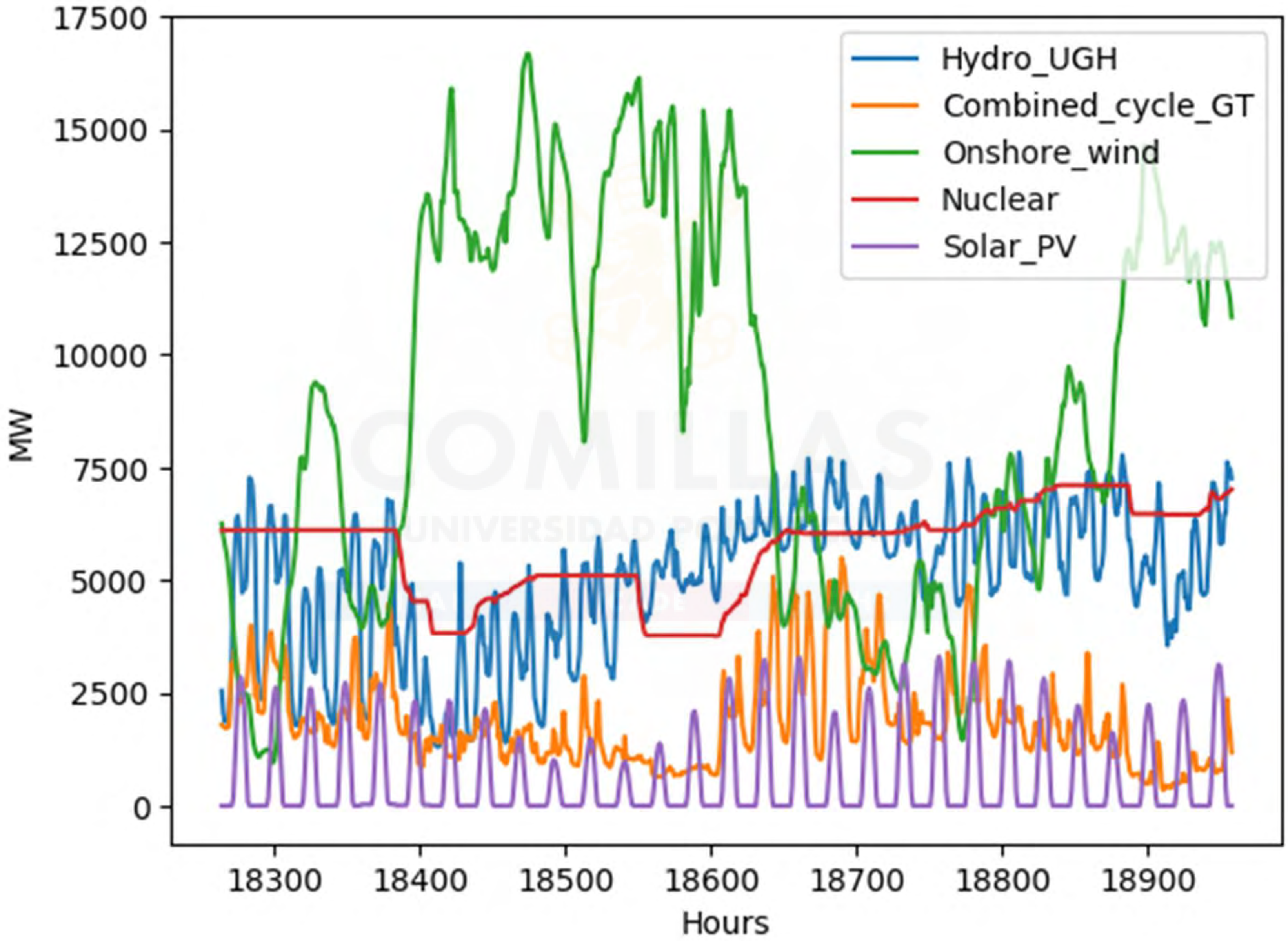
Large hydro, CCGT, onshore wind, nuclear and solar PV output



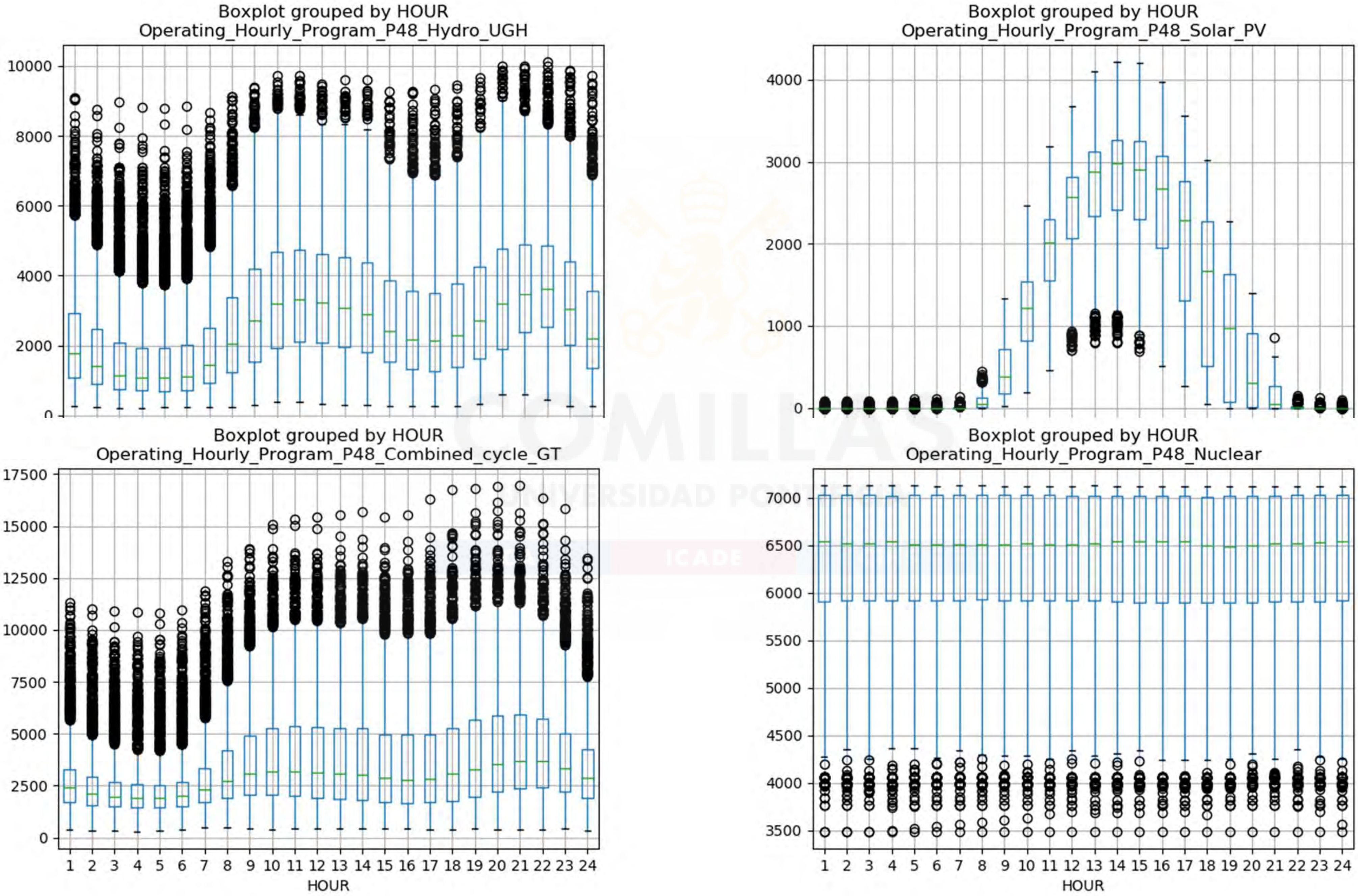
Large hydro, CCGT, onshore wind, nuclear and solar PV output in 2016



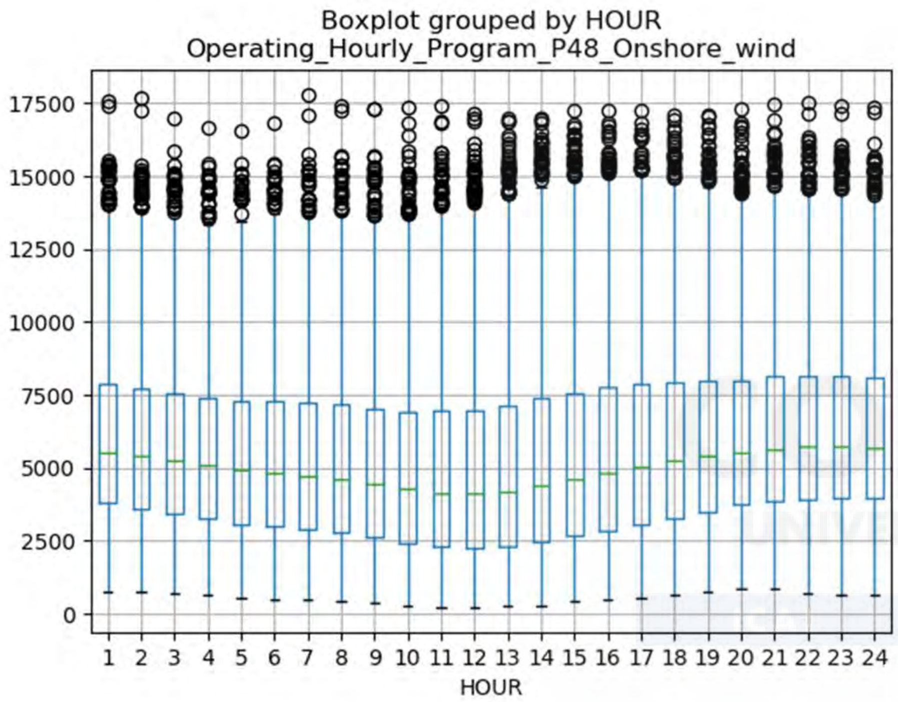
Large hydro, CCGT, onshore wind, nuclear and solar PV output in February 2016

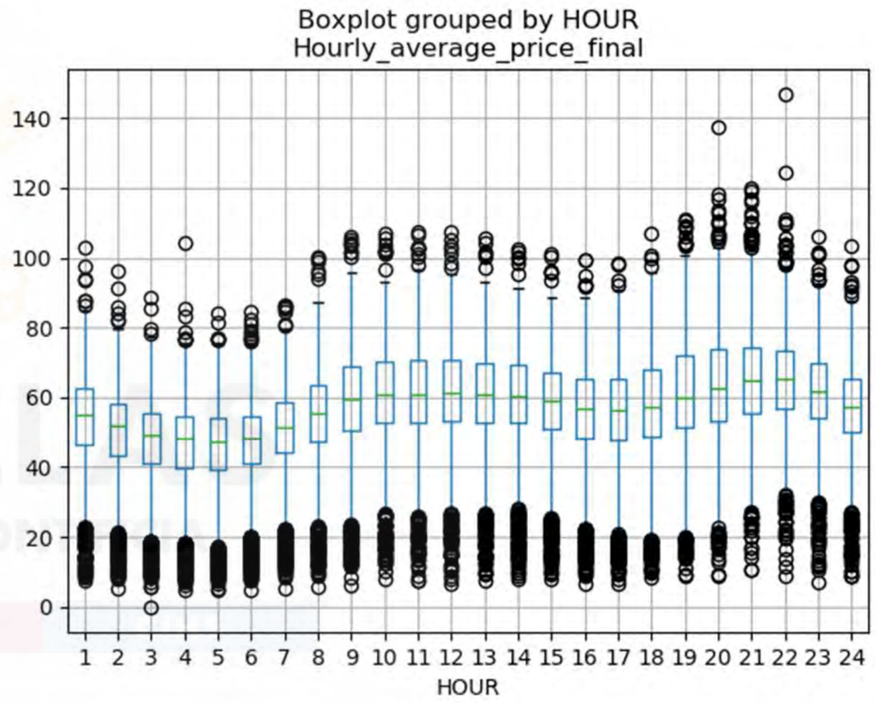


Large hydro, solar PV, CCGT, and nuclear output

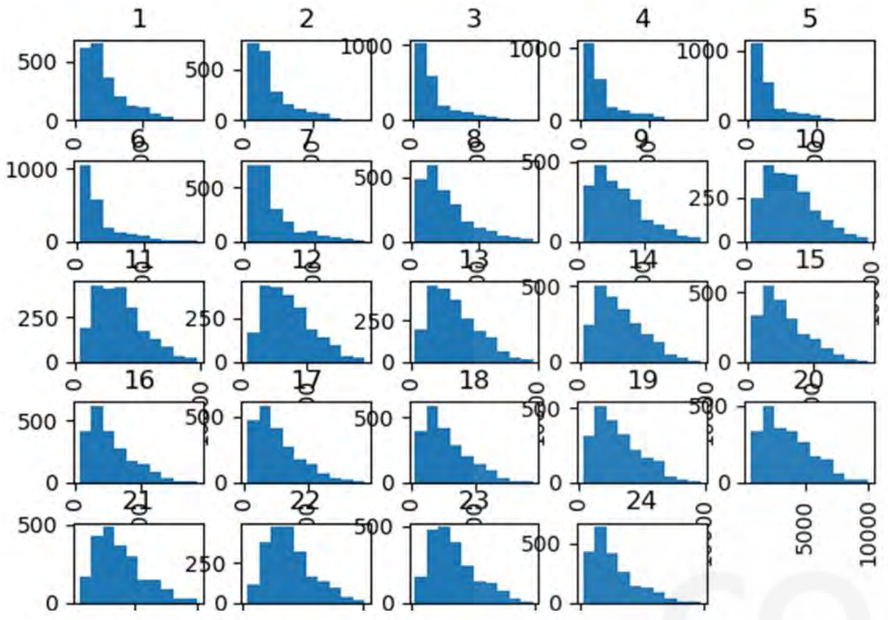


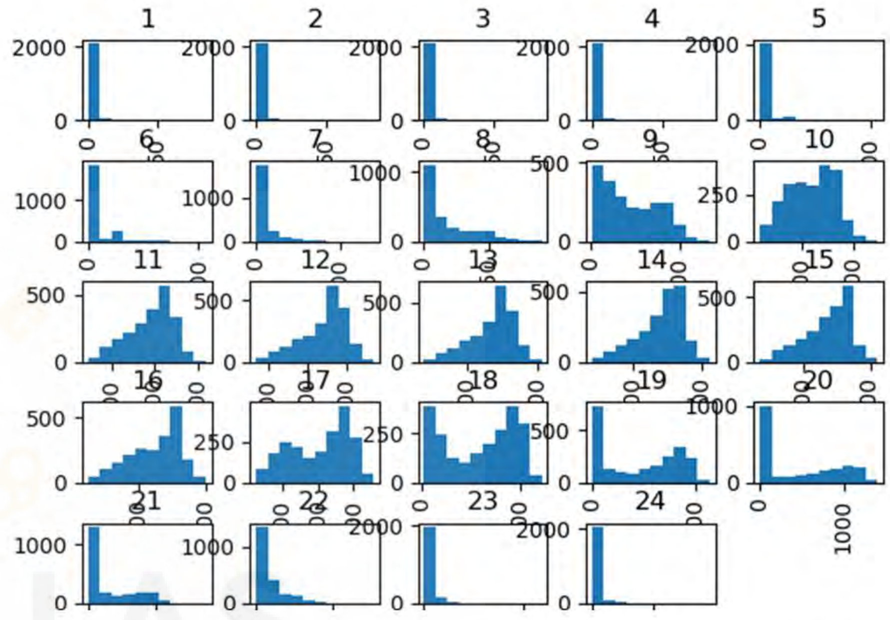
Onshore wind output and price

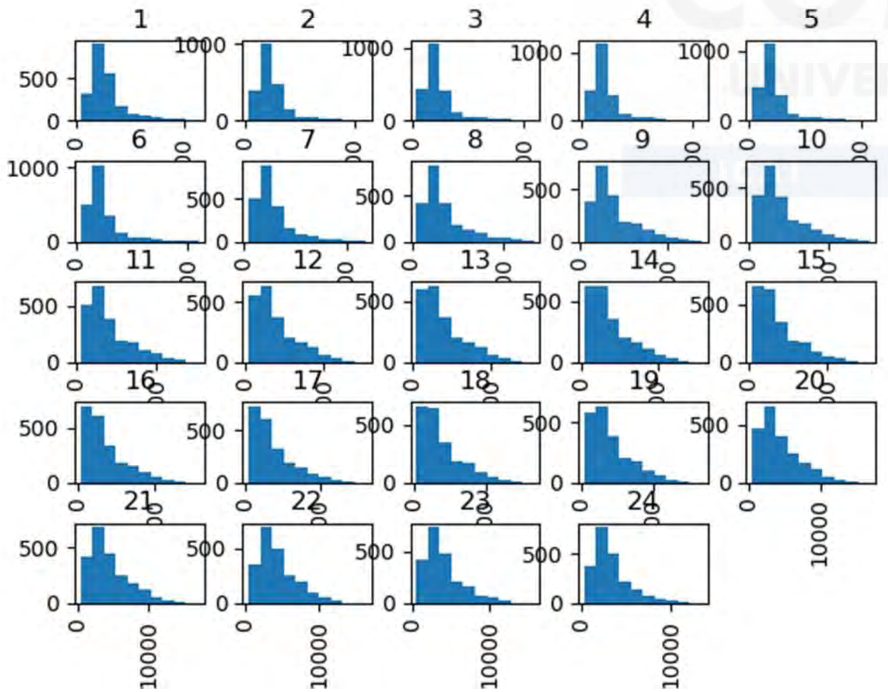


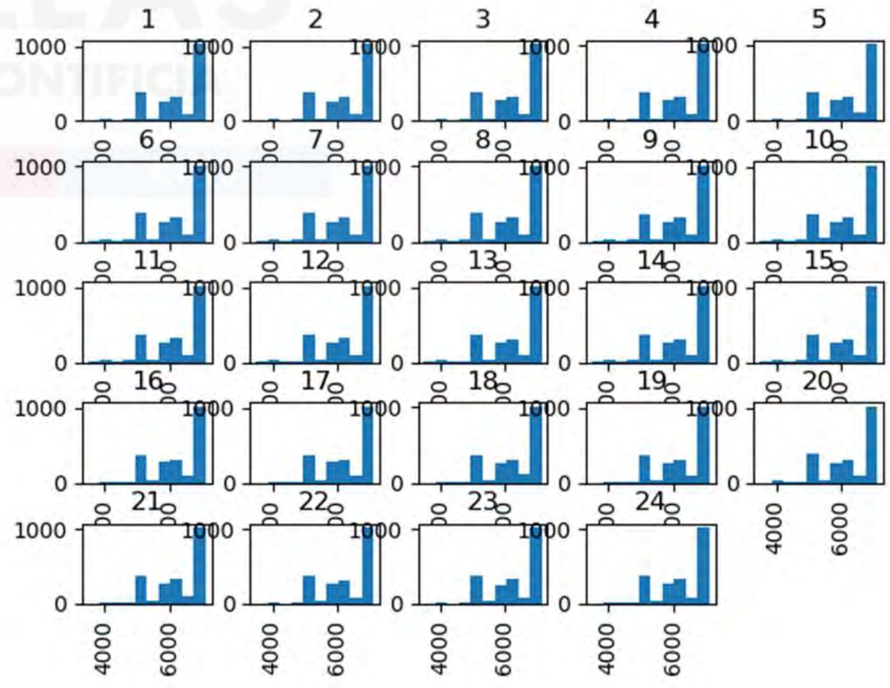


Large hydro, solar PV, CCGT, and nuclear output

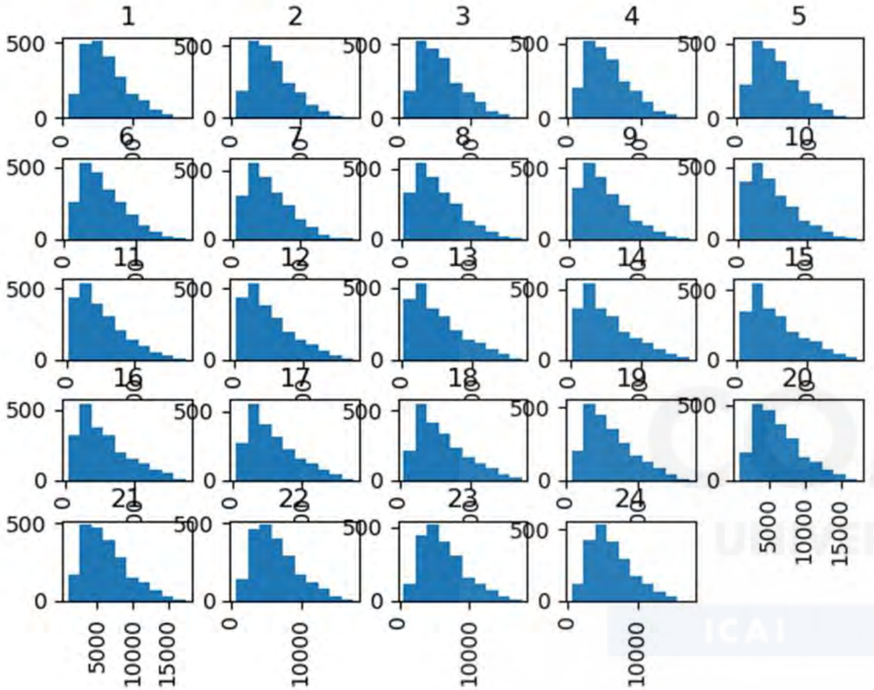


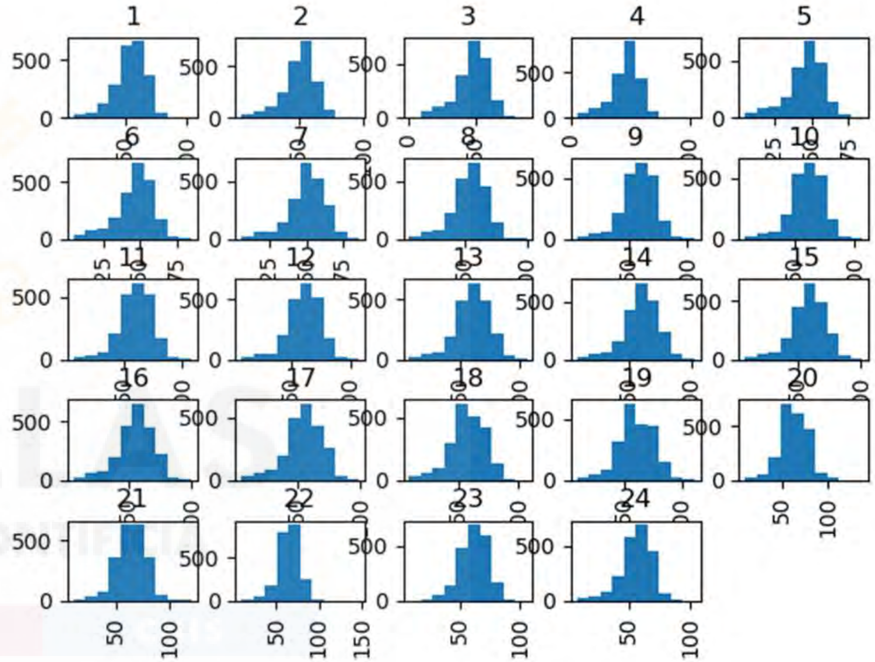






Onshore wind output and price





3

Data analysis. ANOVA

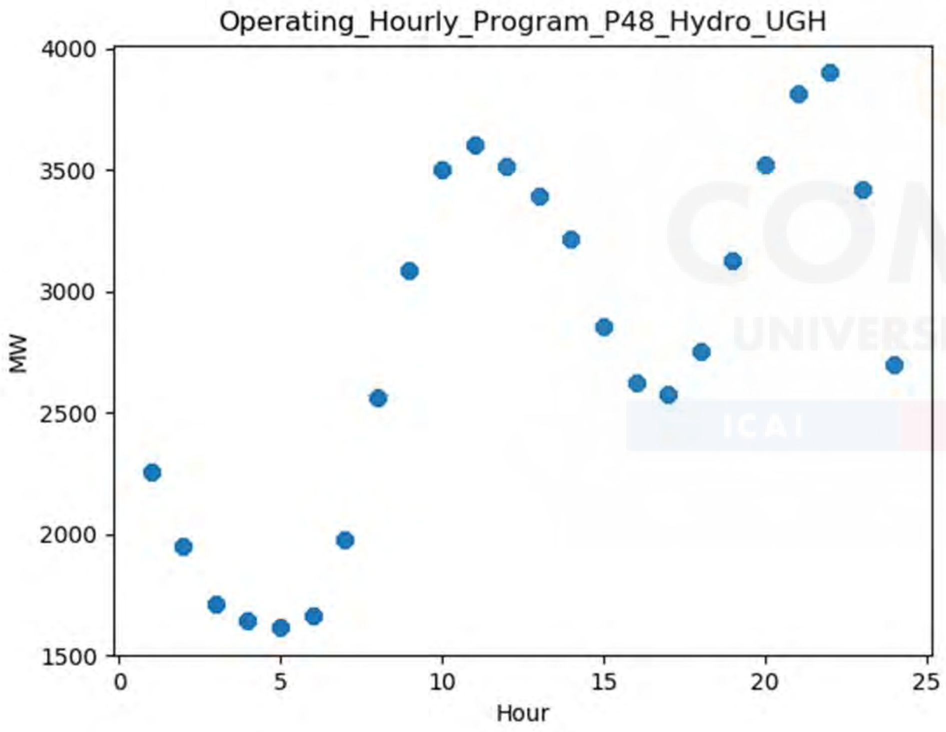
Does the generation of the different technologies vary with factors such as the month, the day of the week, or the hour?

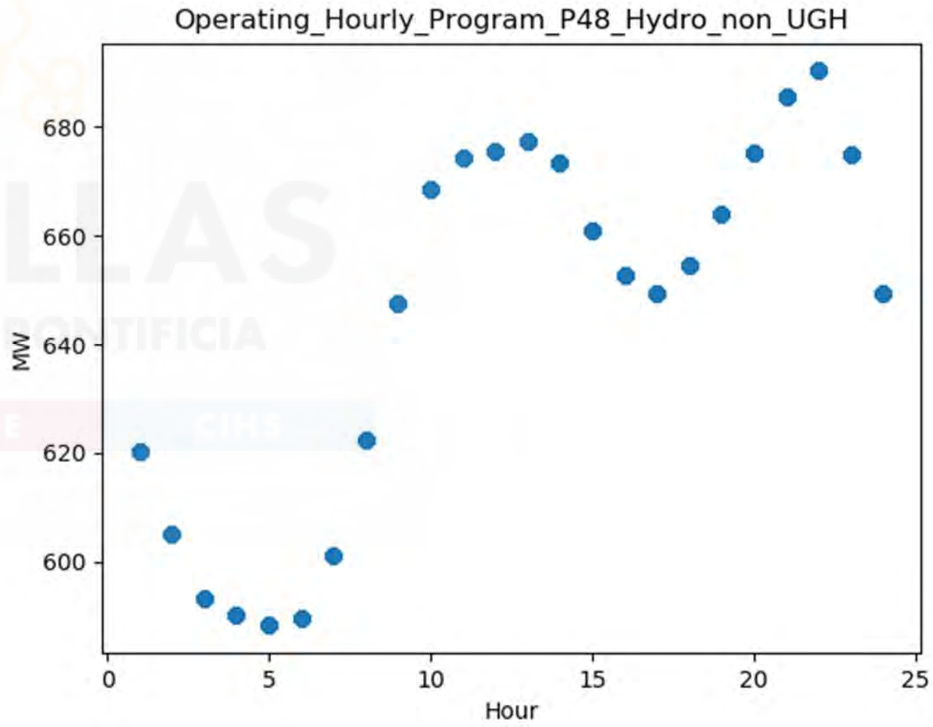
- Which technique can be used?
- What can we expect concerning nuclear output? And CCGT output? And hydro output? And solar PV?

Large and small hydro. One way

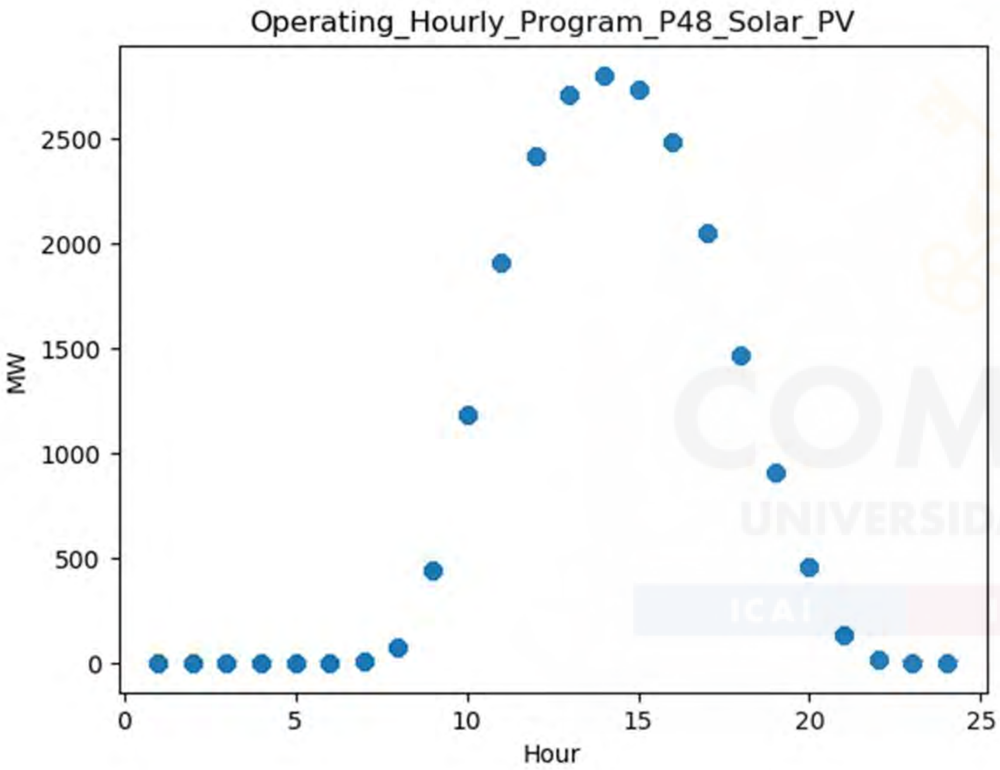
	sum_sq	df	F	PR(>F)
C(HOUR)	2.757722e+10	23.0	402.212312	0.0
Residual	1.566833e+11	52560.0	NaN	NaN

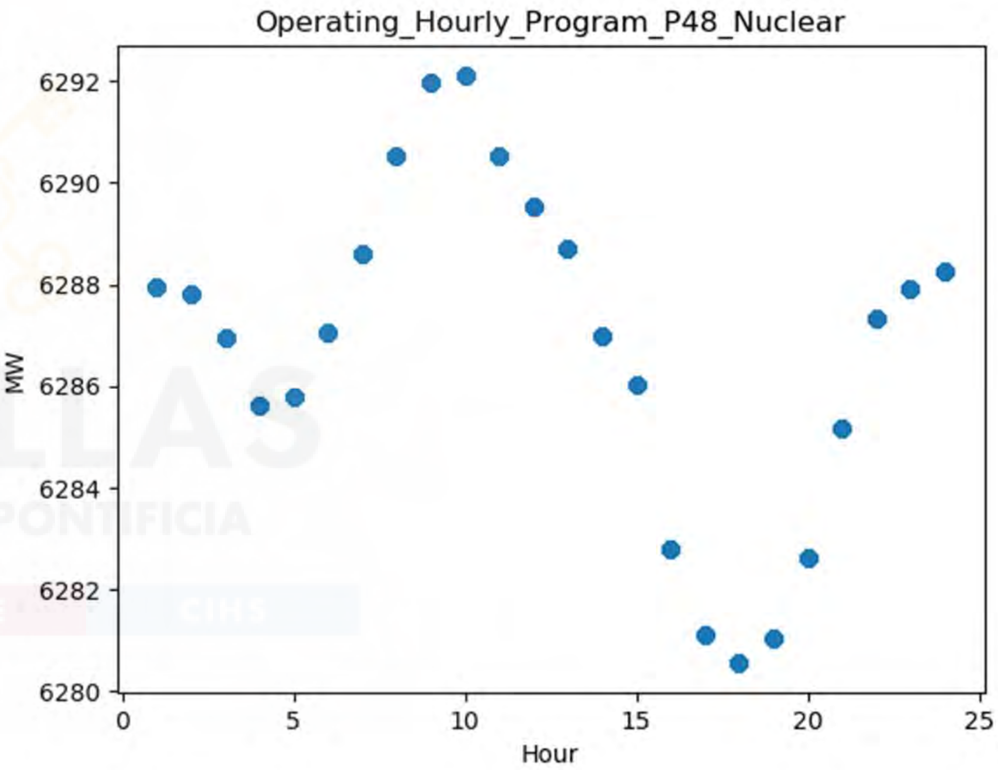
	sum_sq	df	F	PR(>F)
C(HOUR)	5.944270e+07	23.0	37.637572	1.251950e-166
Residual	3.609146e+09	52560.0	NaN	NaN



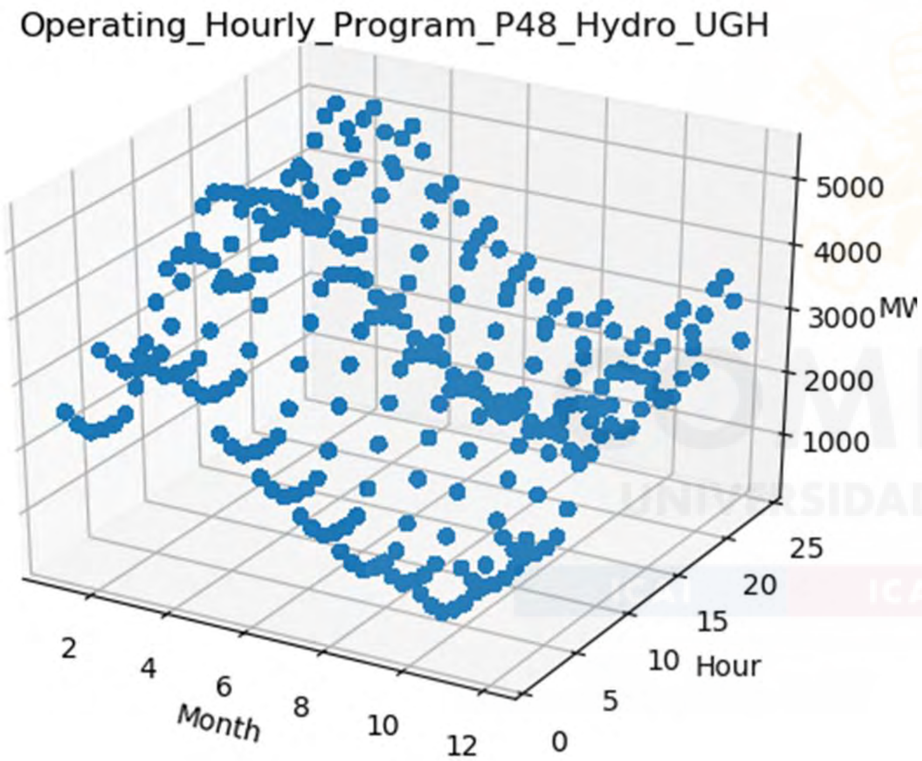


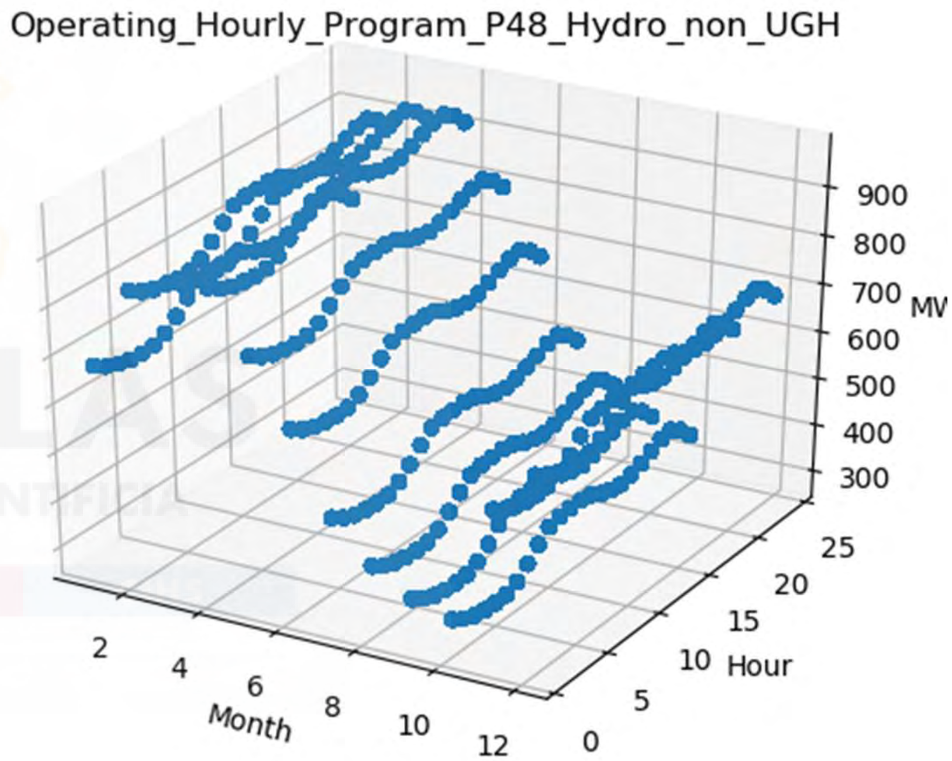
Solar PV and Nuclear. One way





Large and small hydro. Two way

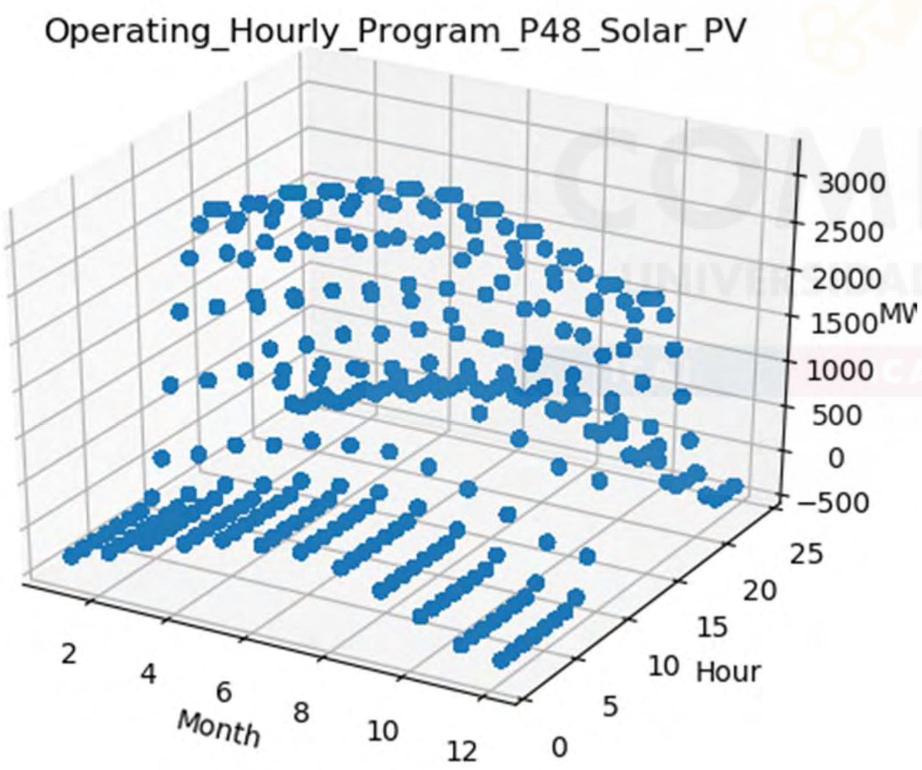


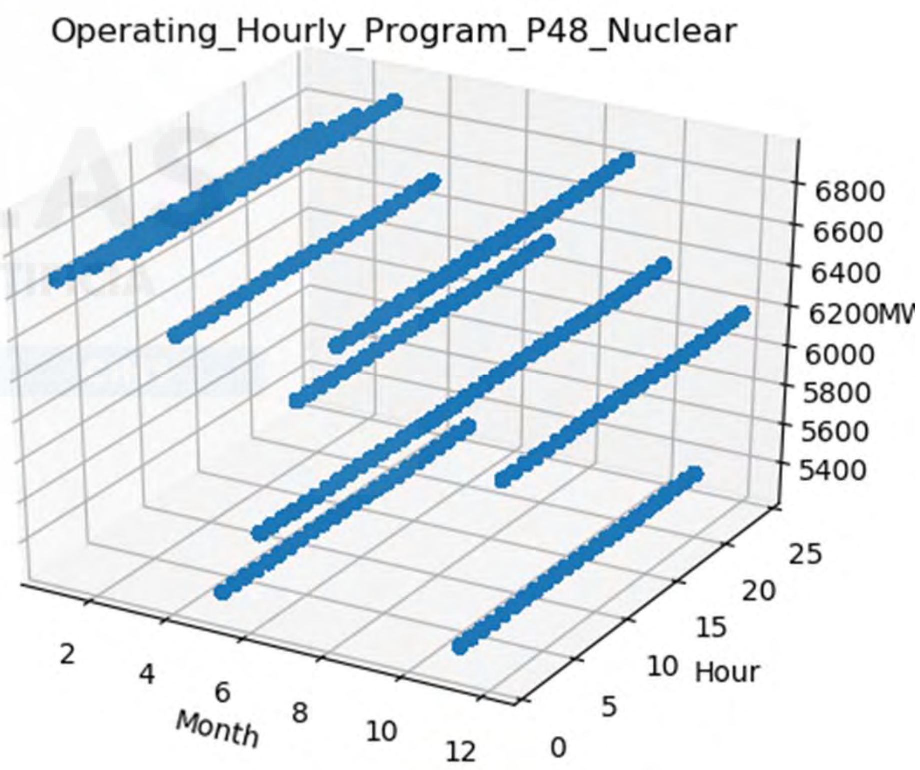


Solar PV and Nuclear. Two way

	sum_sq	df	F	PR(>F)
C(HOUR)	6.032674e+10	23.0	17508.451691	0.0
C(MONTH)	2.973779e+09	11.0	1804.603473	0.0
Residual	7.872247e+09	52549.0	NaN	NaN

	sum_sq	df	F	PR(>F)
C(HOUR)	5.468002e+05	23.0	0.061124	1.0
C(MONTH)	1.536693e+10	11.0	3591.738722	0.0
Residual	2.043874e+10	52549.0	NaN	NaN

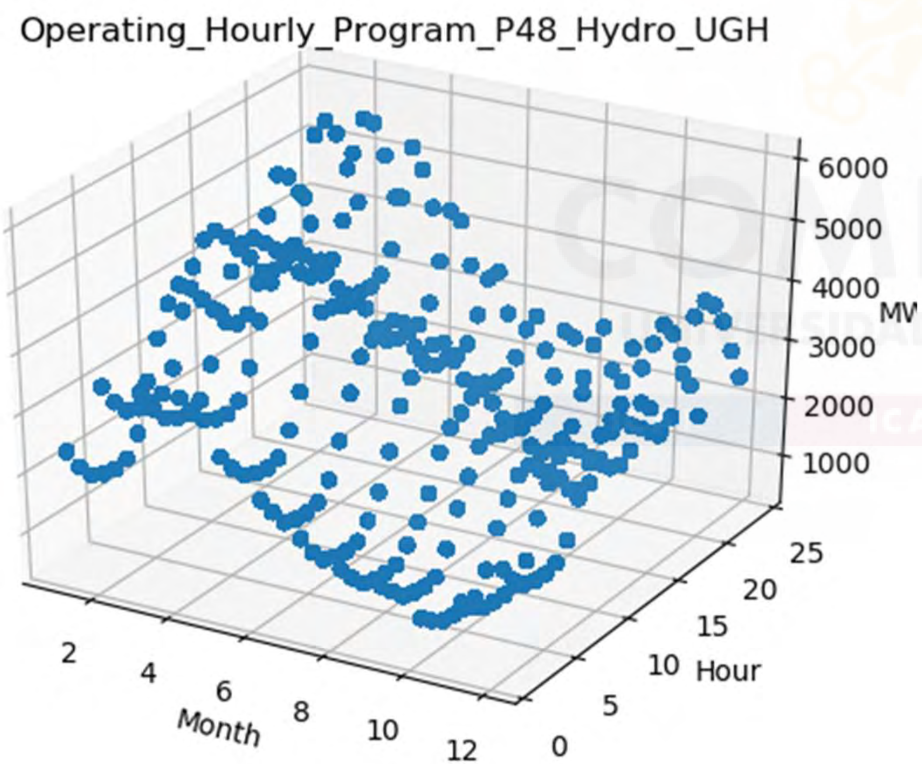


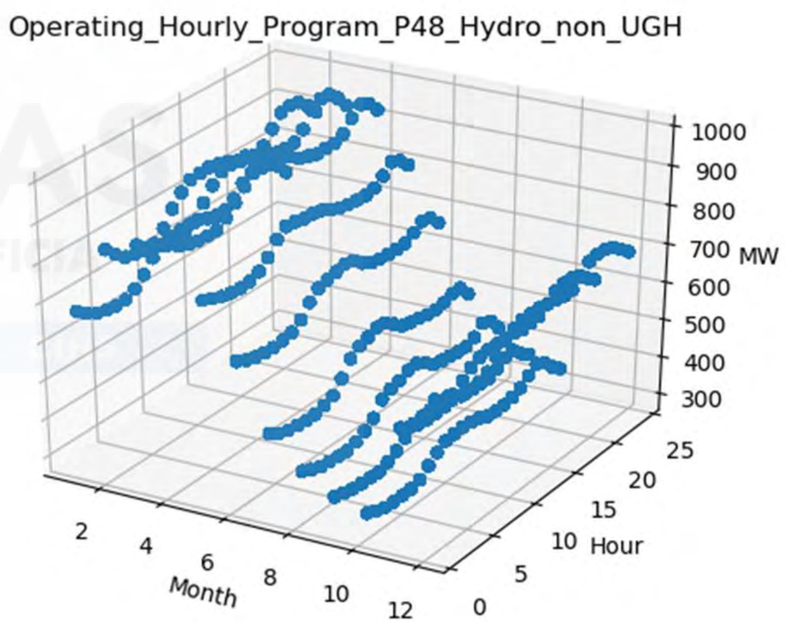


Large and small hydro. Two way with interaction

	sum_sq	df	F	PR(>F)
C(HOUR)	2.754246e+10	23.0	576.082481	0.0
C(MONTH)	4.287686e+10	11.0	1875.167349	0.0
C(HOUR):C(MONTH)	5.099106e+09	253.0	9.695790	0.0
Residual	1.087073e+11	52296.0	NaN	NaN

	sum_sq	df	F	PR(>F)
C(HOUR)	5.909998e+07	23.0	83.811830	0.000000
C(MONTH)	1.997213e+09	11.0	5922.125213	0.000000
C(HOUR):C(MONTH)	8.603591e+06	253.0	1.109188	0.112736
Residual	1.603329e+09	52296.0	NaN	NaN

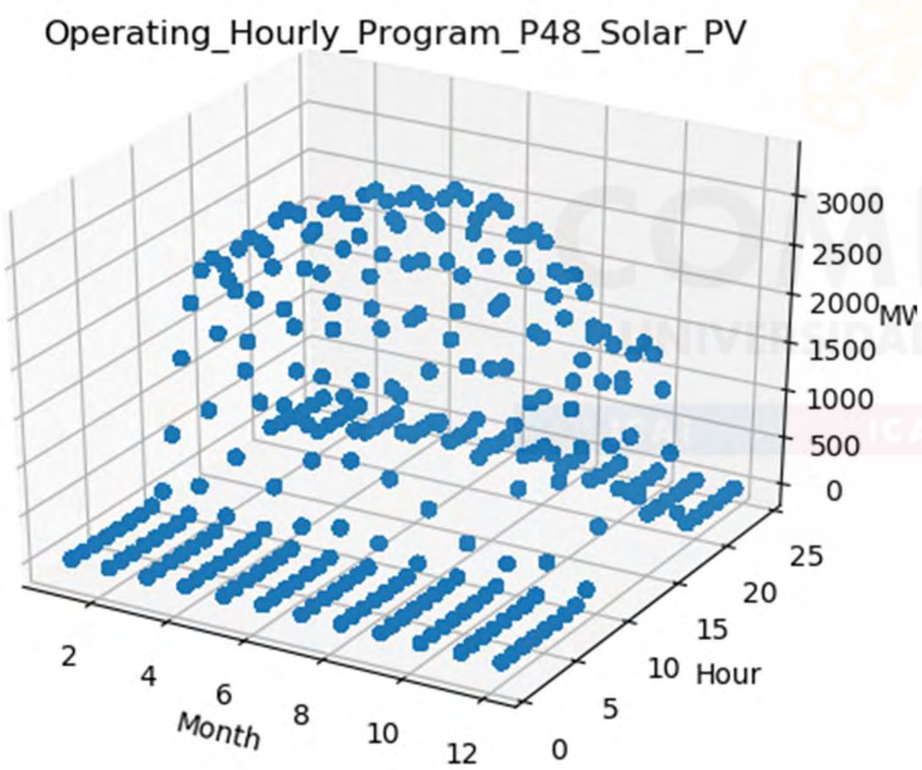


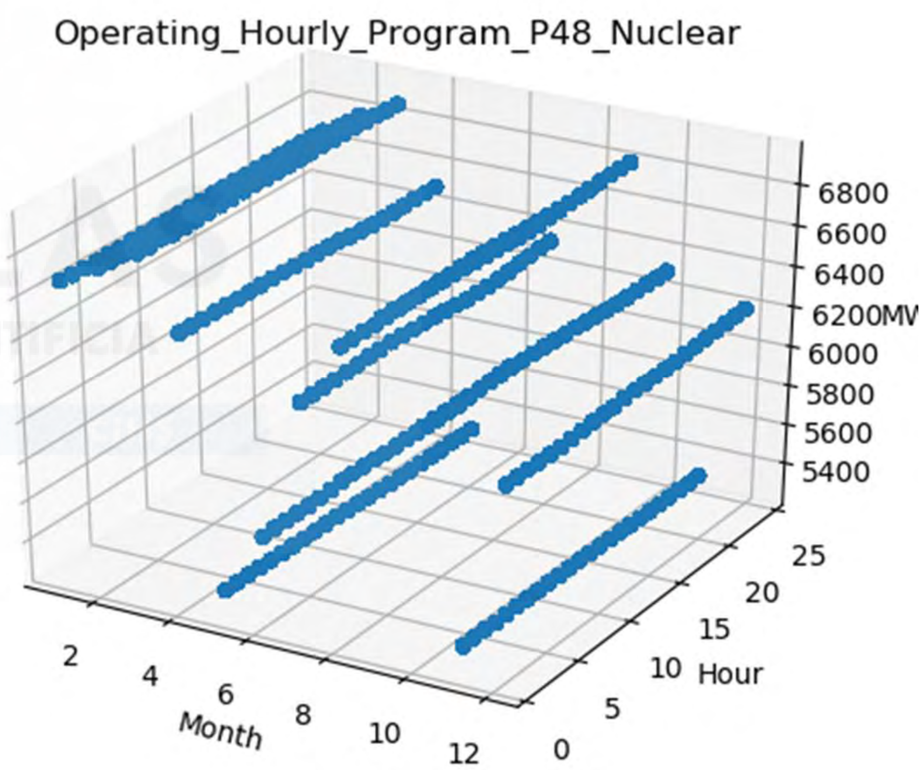


Solar PV and Nuclear. Two way with interaction

	sum_sq	df	F	PR(>F)
C(HOUR)	6.032674e+10	23.0	33427.604221	0.0
C(MONTH)	2.973779e+09	11.0	3445.397213	0.0
C(HOUR):C(MONTH)	3.768834e+09	253.0	189.849629	0.0
Residual	4.103413e+09	52296.0	NaN	NaN

	sum_sq	df	F	PR(>F)
C(HOUR)	5.468002e+05	23.0	0.060843	1.0
C(MONTH)	1.536693e+10	11.0	3575.254680	0.0
C(HOUR):C(MONTH)	4.622416e+06	253.0	0.046759	1.0
Residual	2.043412e+10	52296.0	NaN	NaN





4

Reducing dimensions

COMILLAS UNIVERSIDAD PONTIFICIA ICAI

Principal Components Analysis (PCA) Introduction

- PCA aims to determine a few linear combinations of the original variables that can be used to summarize the data set without losing much information
- PCA refers to the process by which principal components are computed and the subsequent use of these components in understanding the data

When faced with a large set of correlated variables, principal components allow us to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the original set.

Principal Components Analysis (PCA)
What is PCA for?

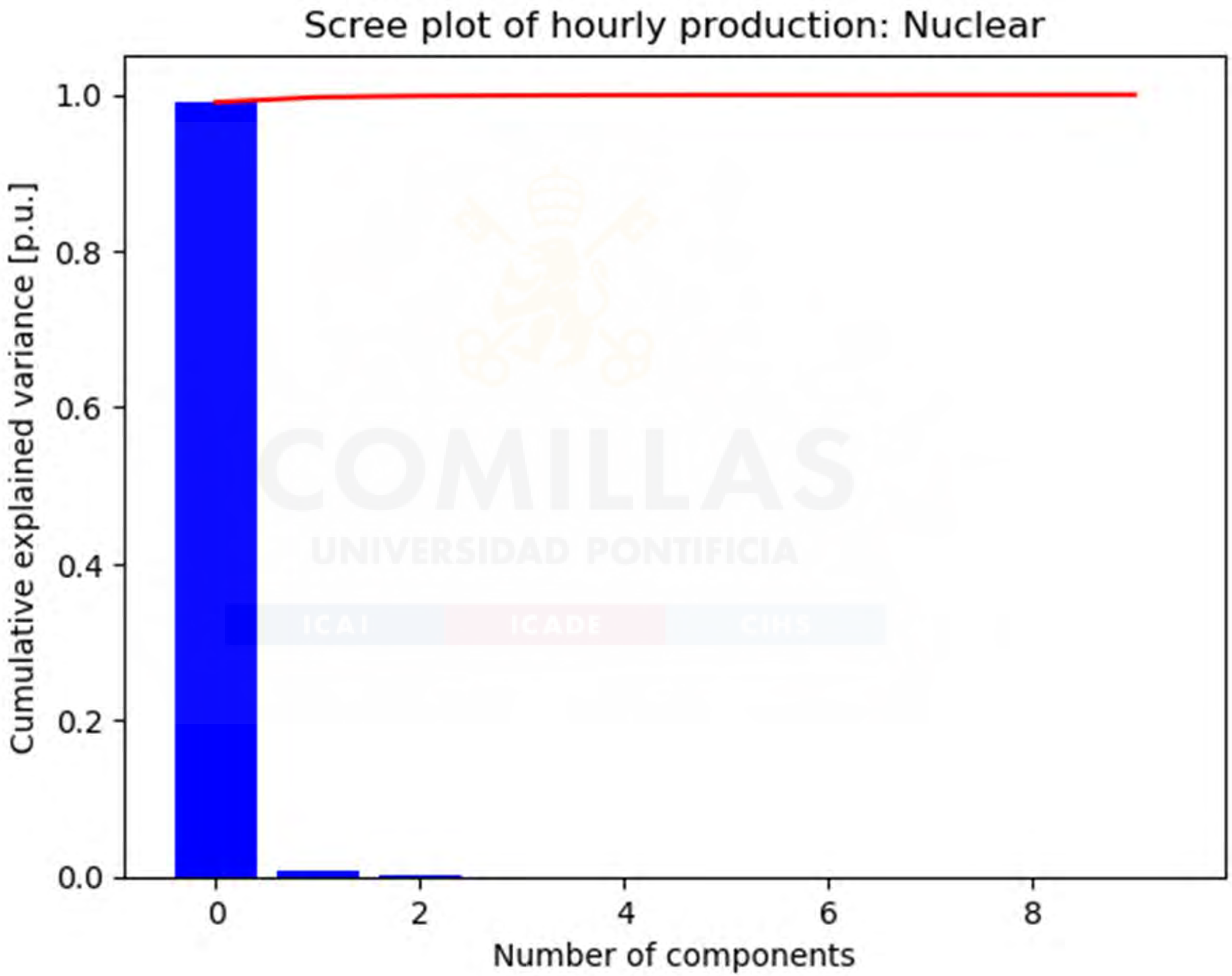
- PCA is useful for finding out explanatory variables of data that are not directly observed
- This technique is used for Regression, Clustering, and Forecasting when the number of input variables is high and/or variables are correlated
- A way of identifying patterns (driving forces) in data
- After these patterns are found, data can be compressed, reducing the number of dimensions without losing much information (variability)

Hourly output for each technology

- Can we reduce the number of hours to a few components?
- How many principal components are reasonable for each technology
- Two components are good enough for storing the original information

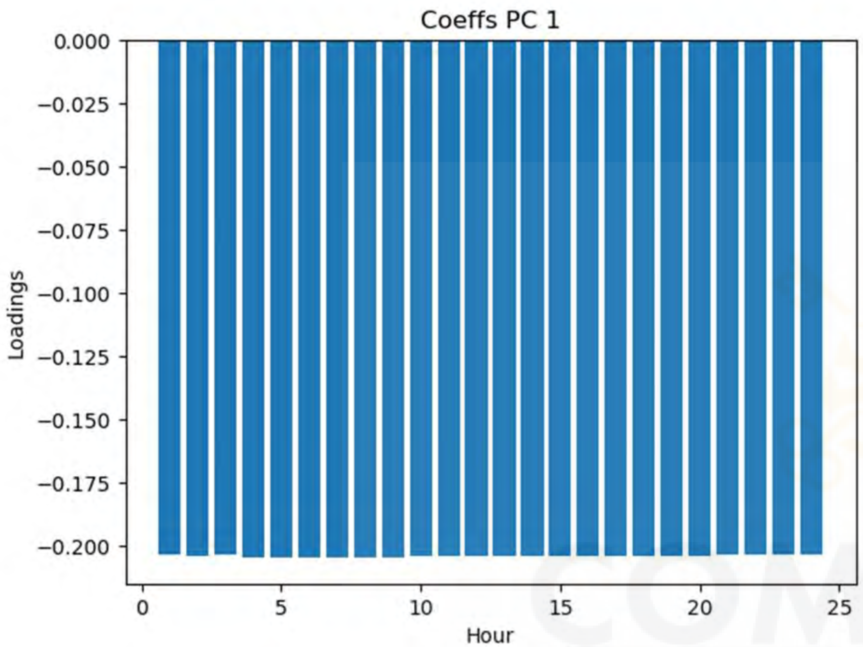
Nuclear

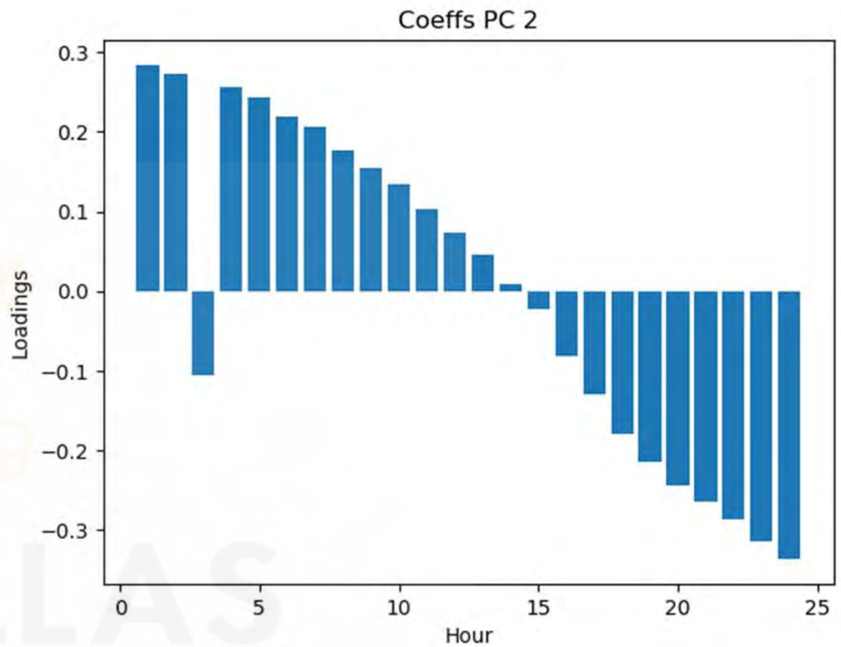
Principal Component Analysis (PCA)

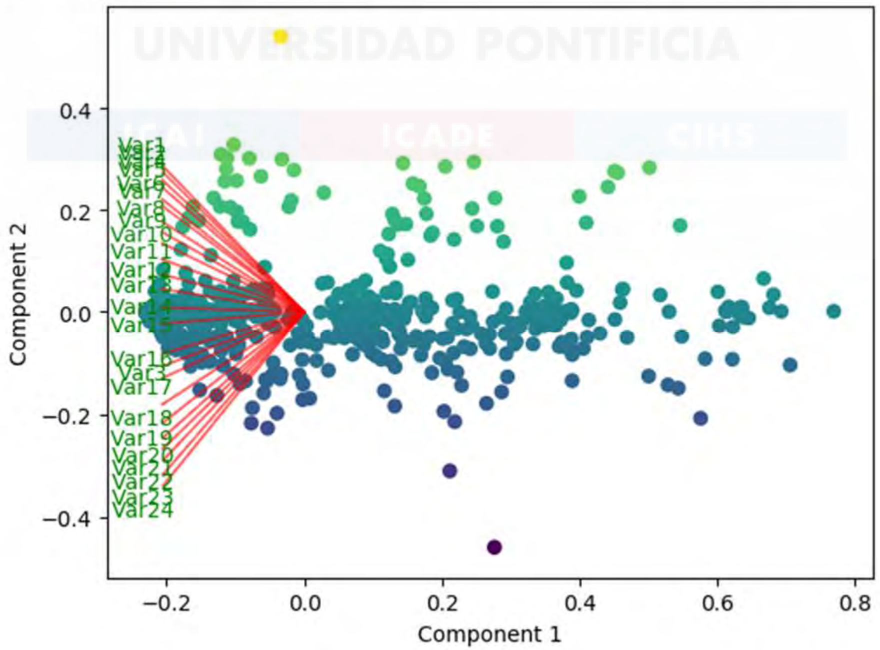


Nuclear

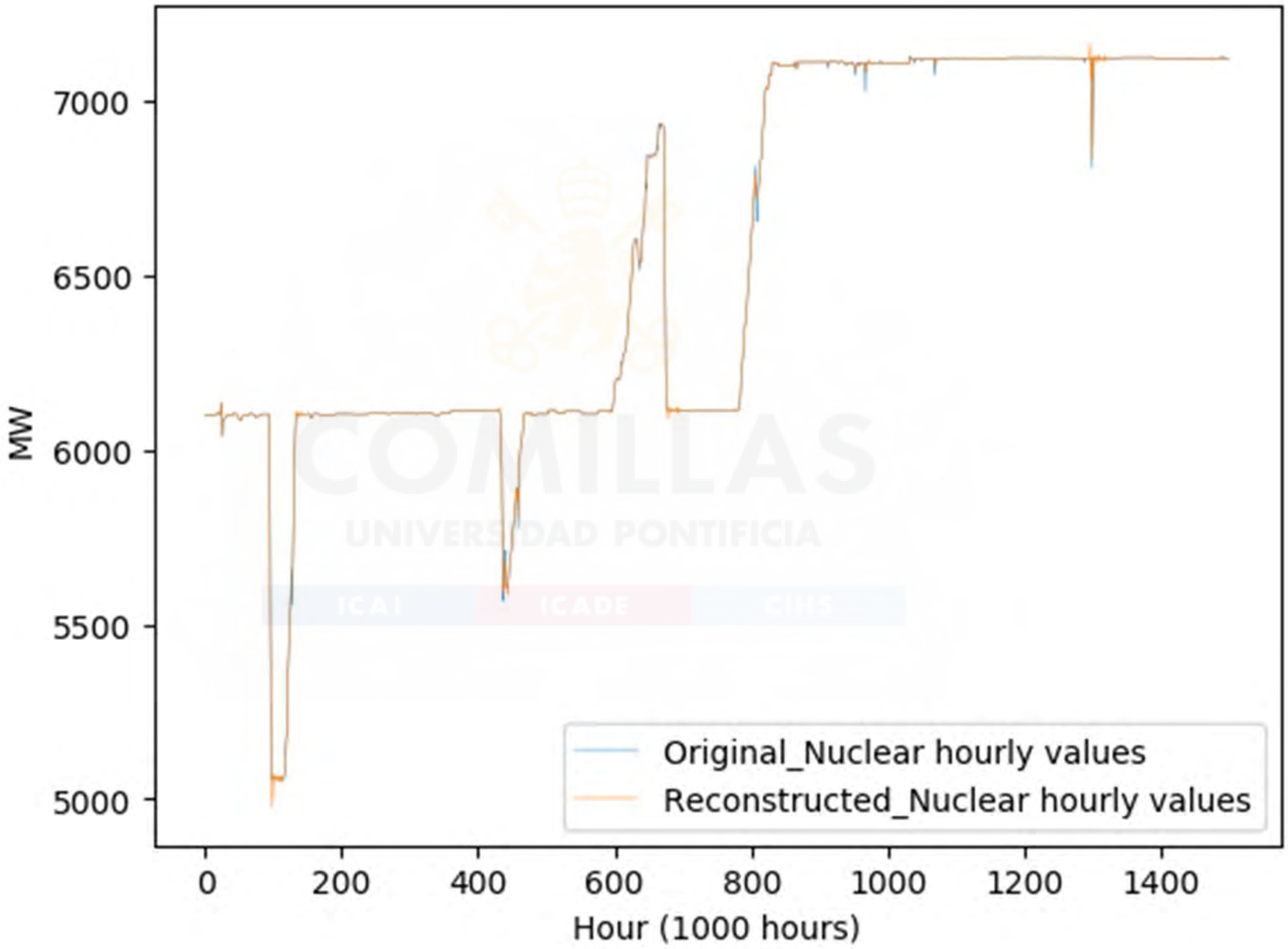
Coefficients of PC1 and PC2 and scores



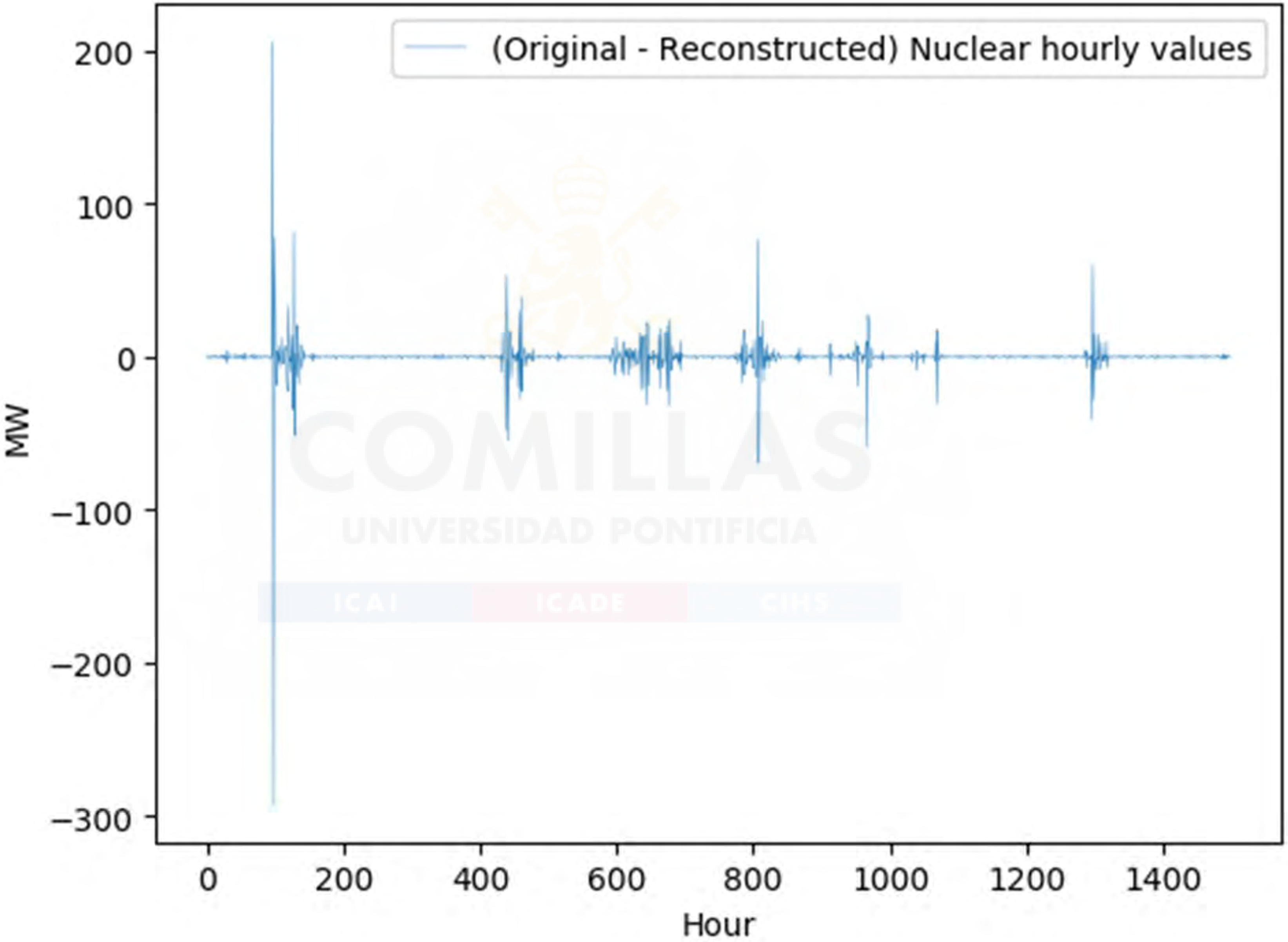




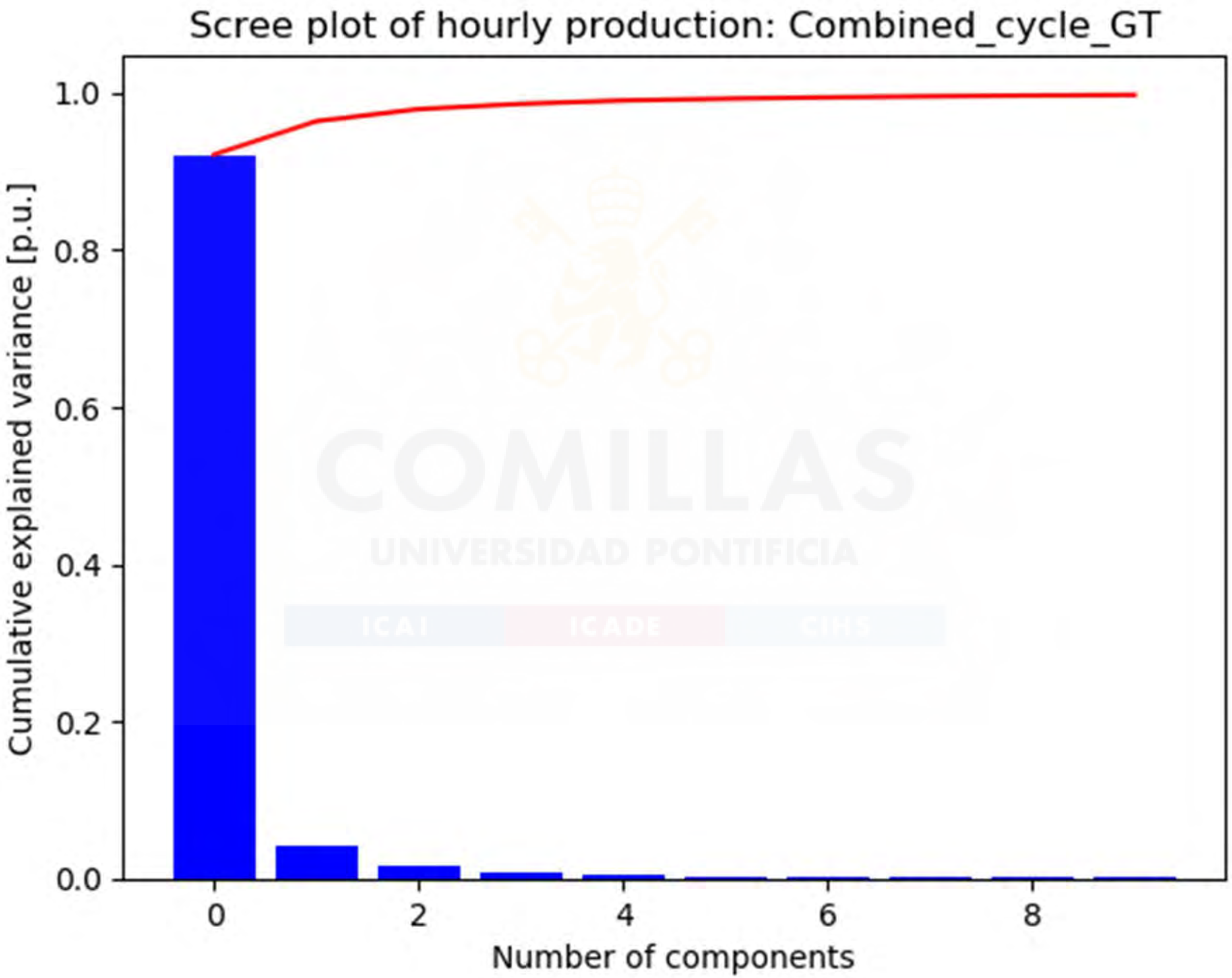
Nuclear
Original and reconstructed hourly values with 2 PCs



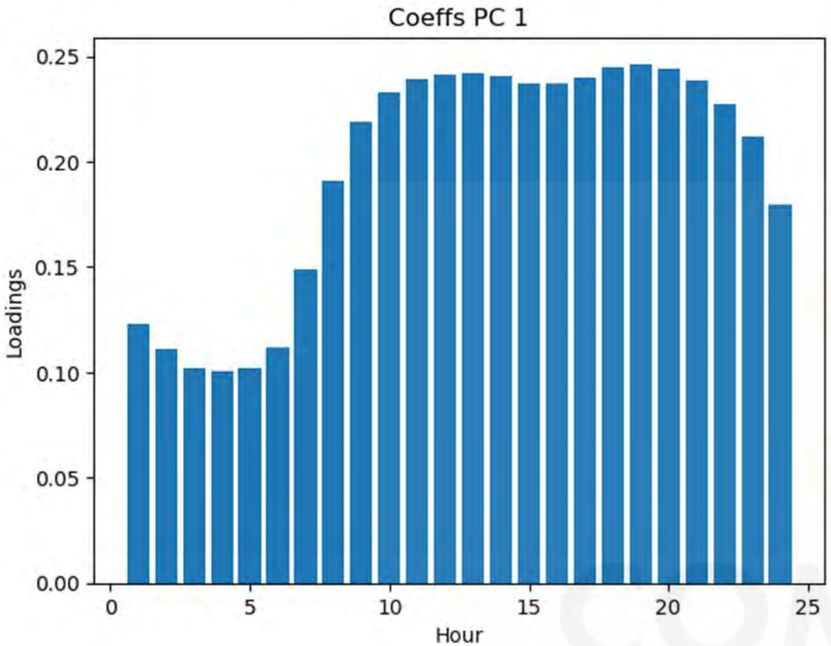
Nuclear
(Original - Reconstructed) hourly values

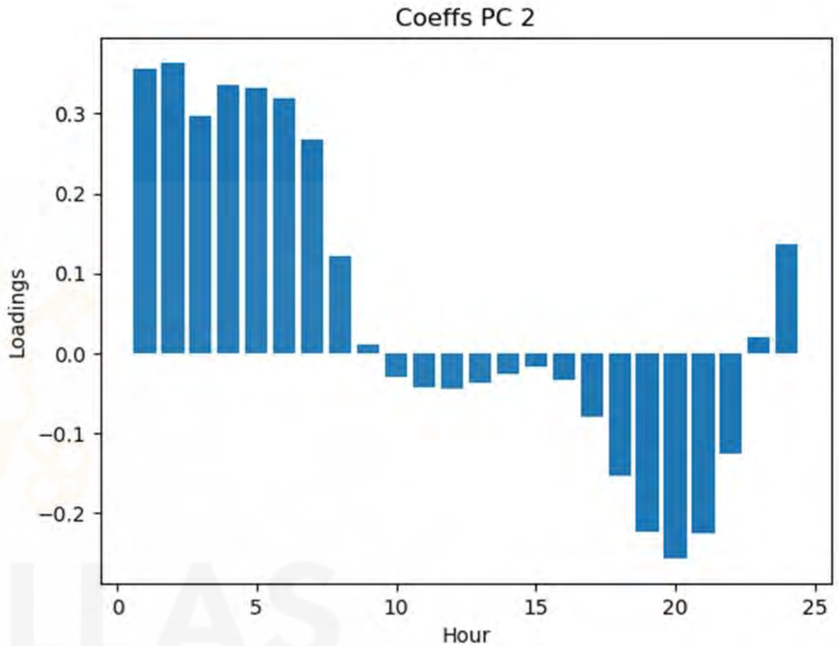


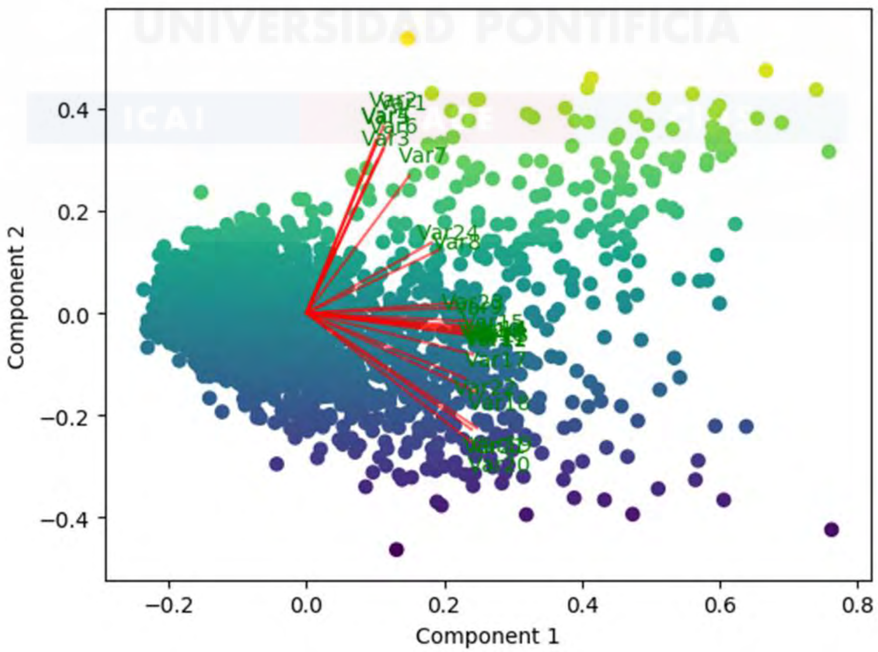
Combined Cycle GT
Principal Component Analysis (PCA)



Combined Cycle GT
Coefficients of PC1 and PC2 and scores

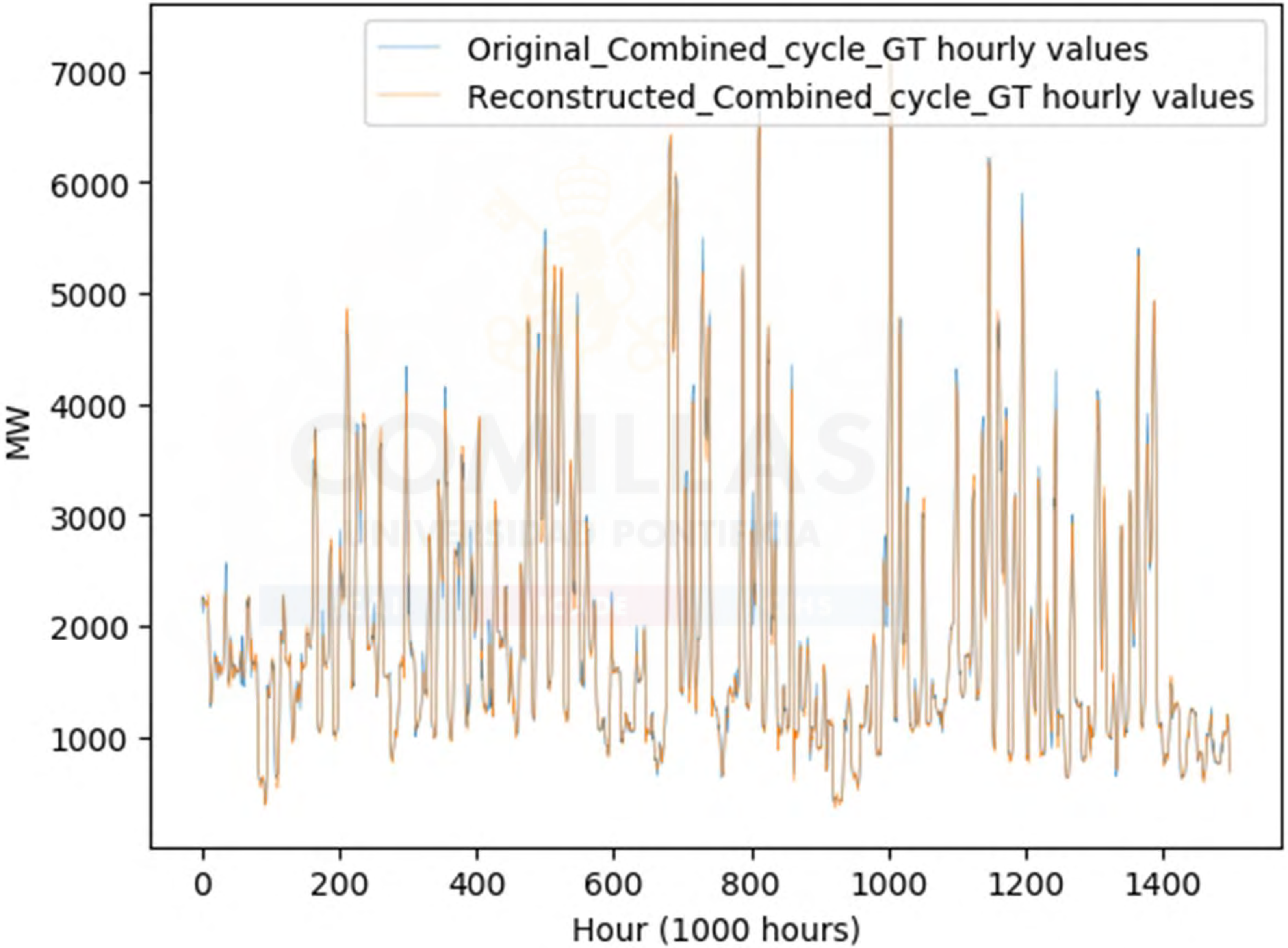




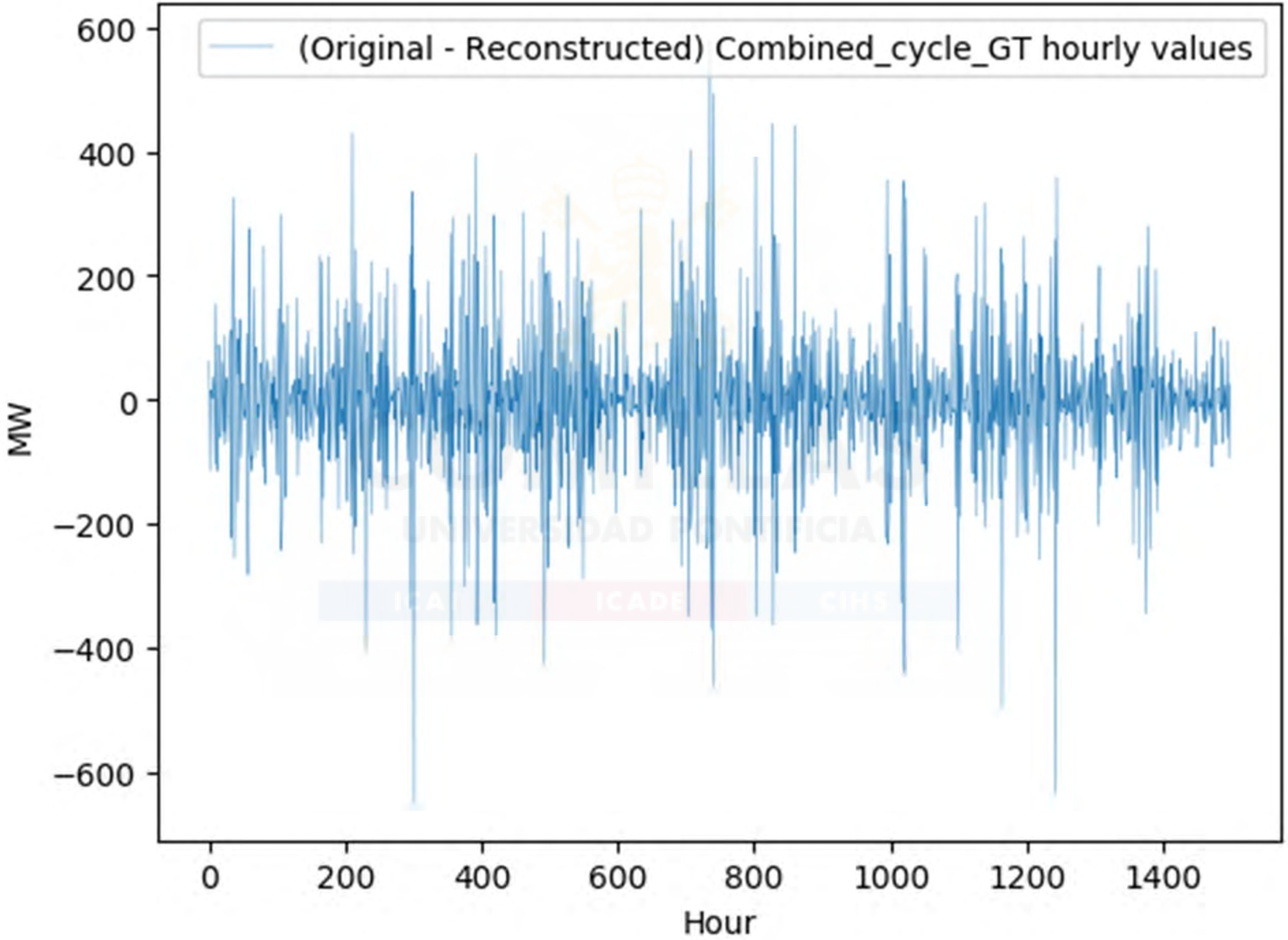


Combined Cycle GT

Original and reconstructed hourly values with 2 PCs

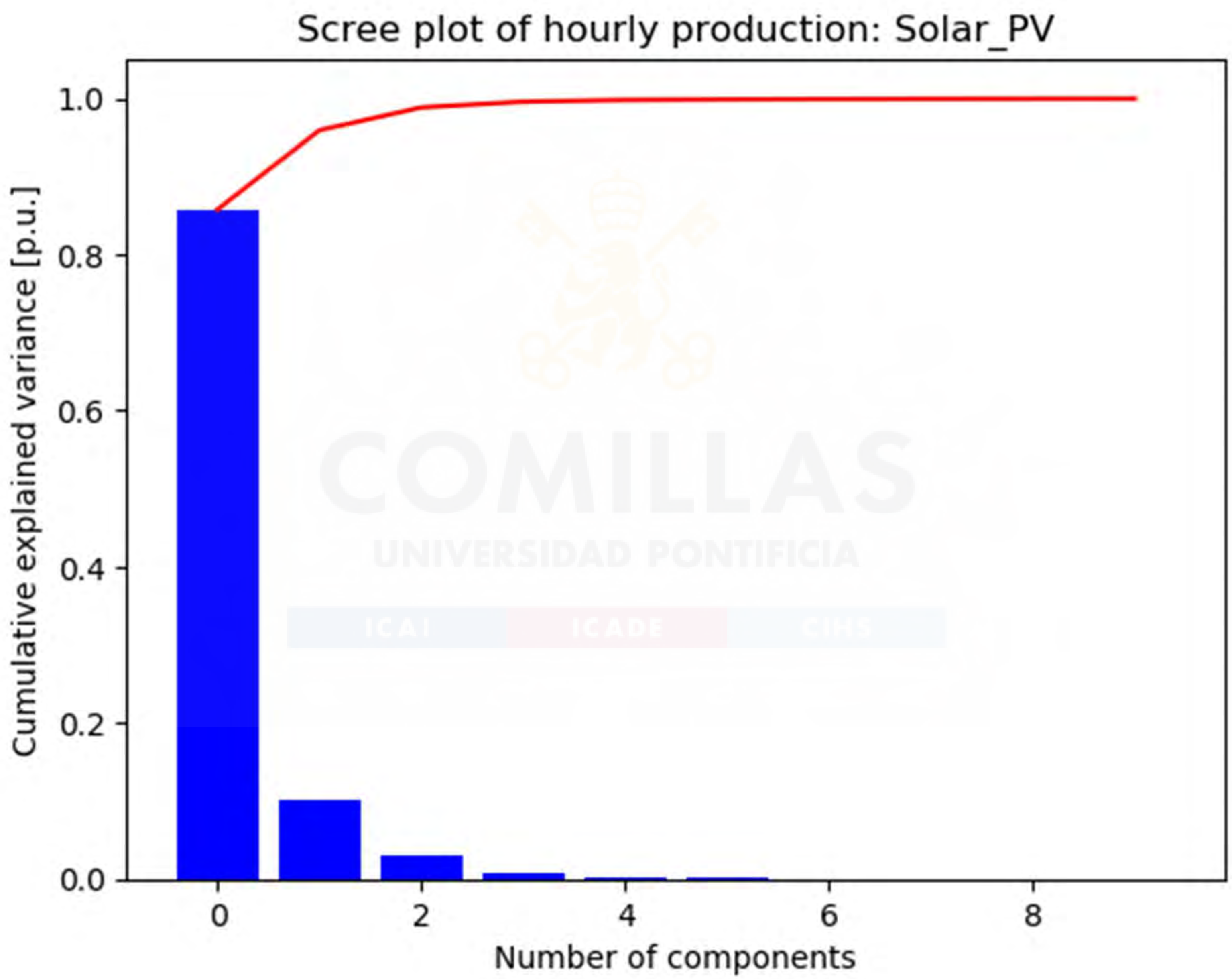


Combined Cycle GT
(Original - Reconstructed) hourly values



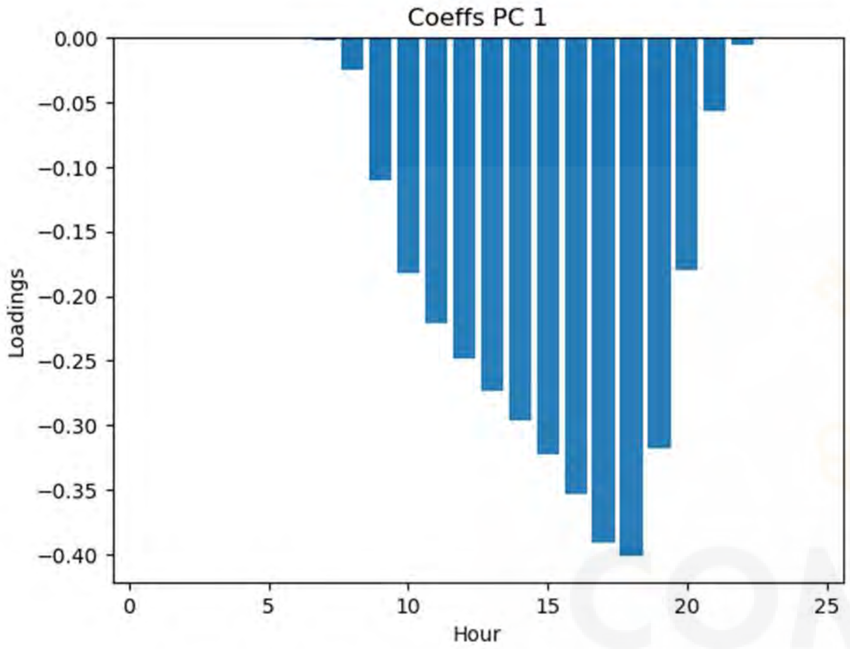
Solar PV

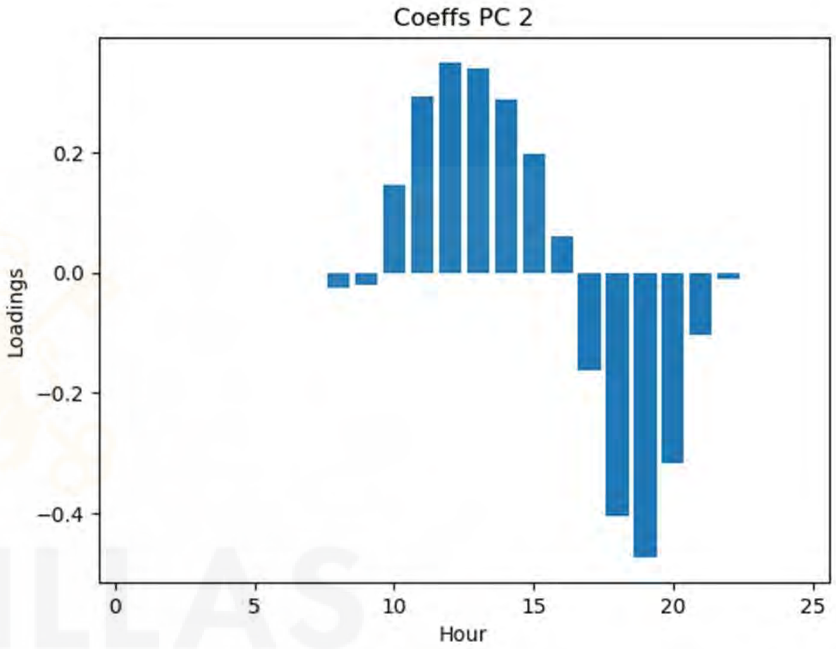
Principal Component Analysis (PCA)

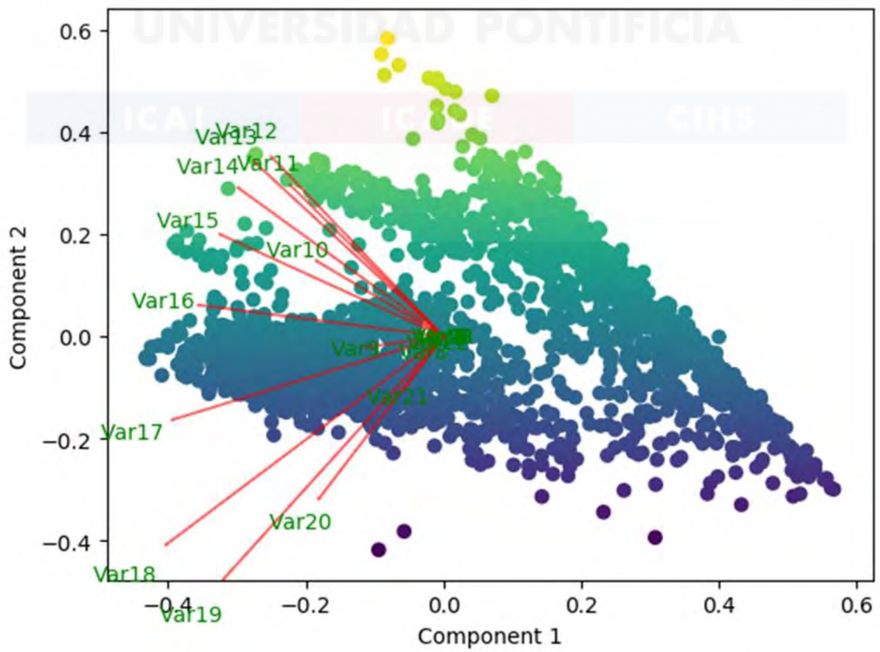


Solar PV

Coefficients of PC1 and PC2 and scores

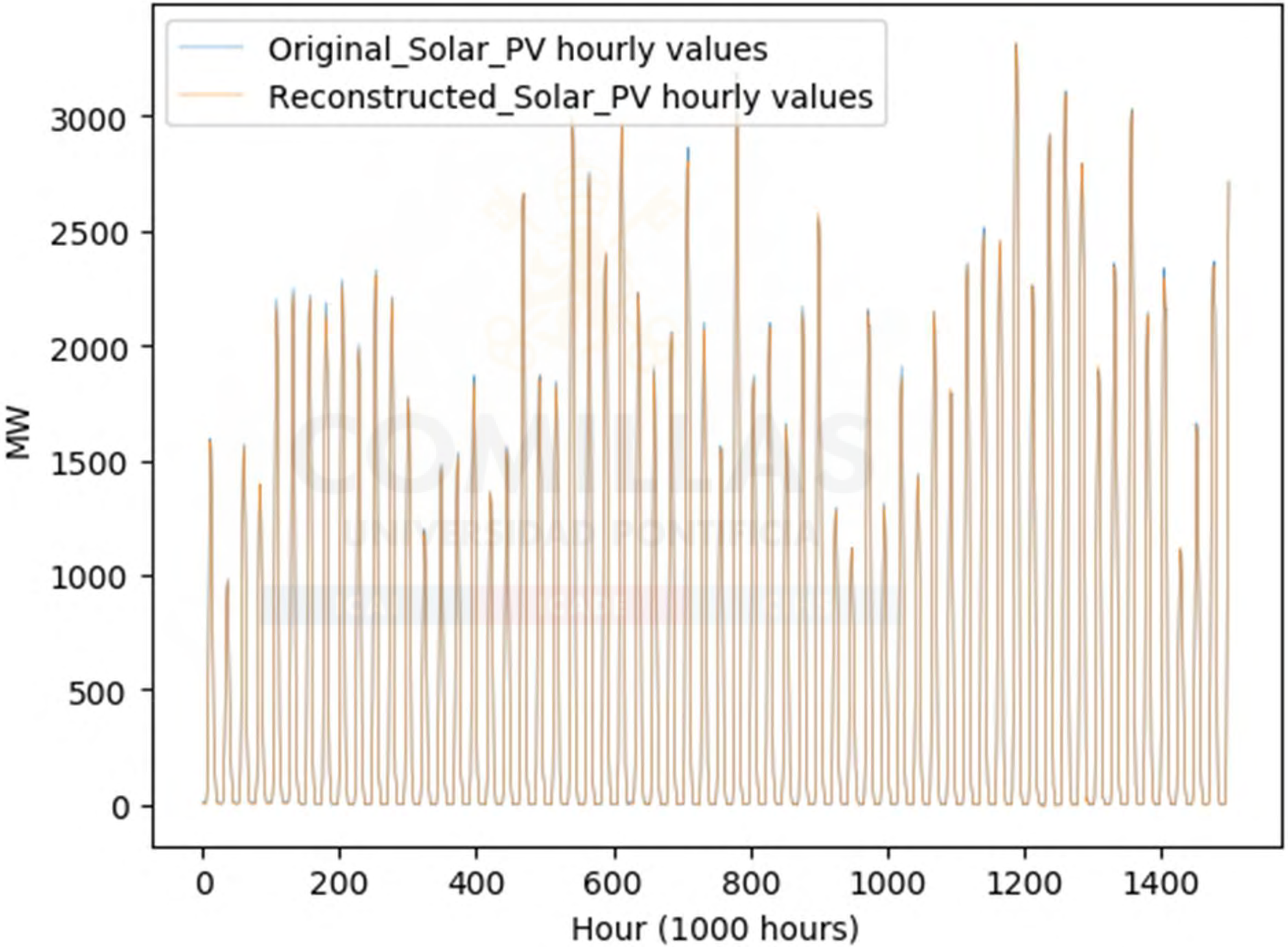




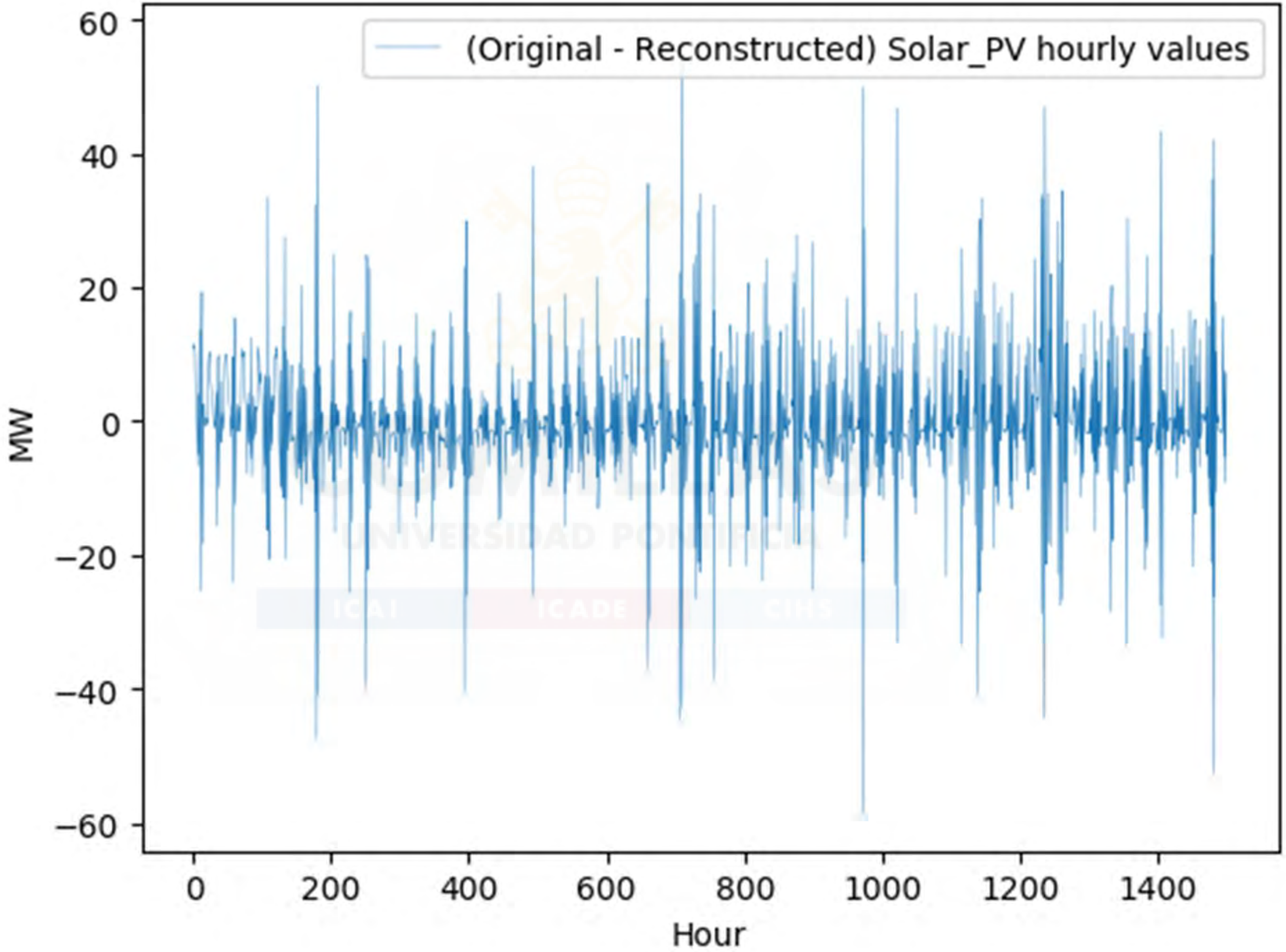


Solar PV

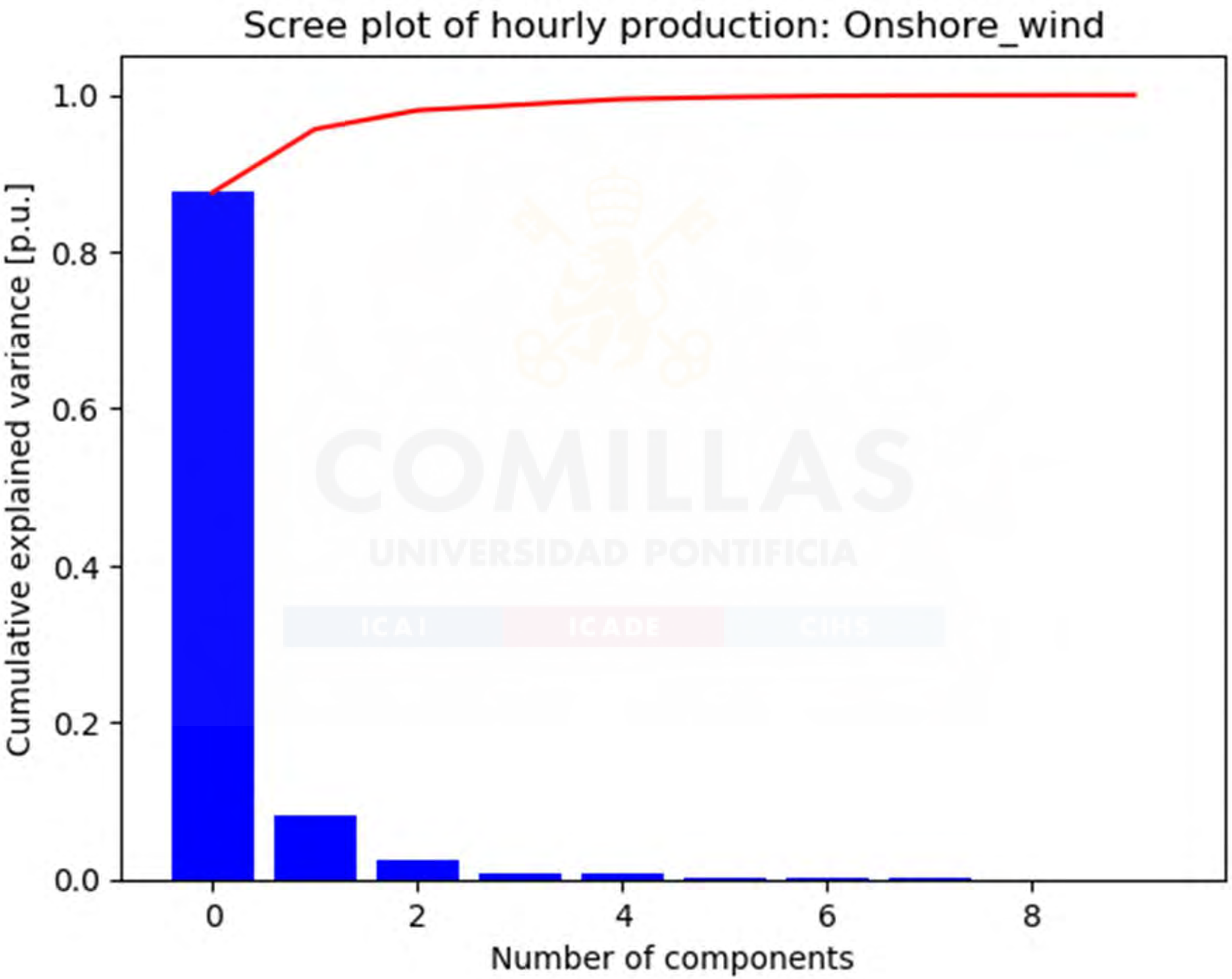
Original and reconstructed hourly values with 2 PCs



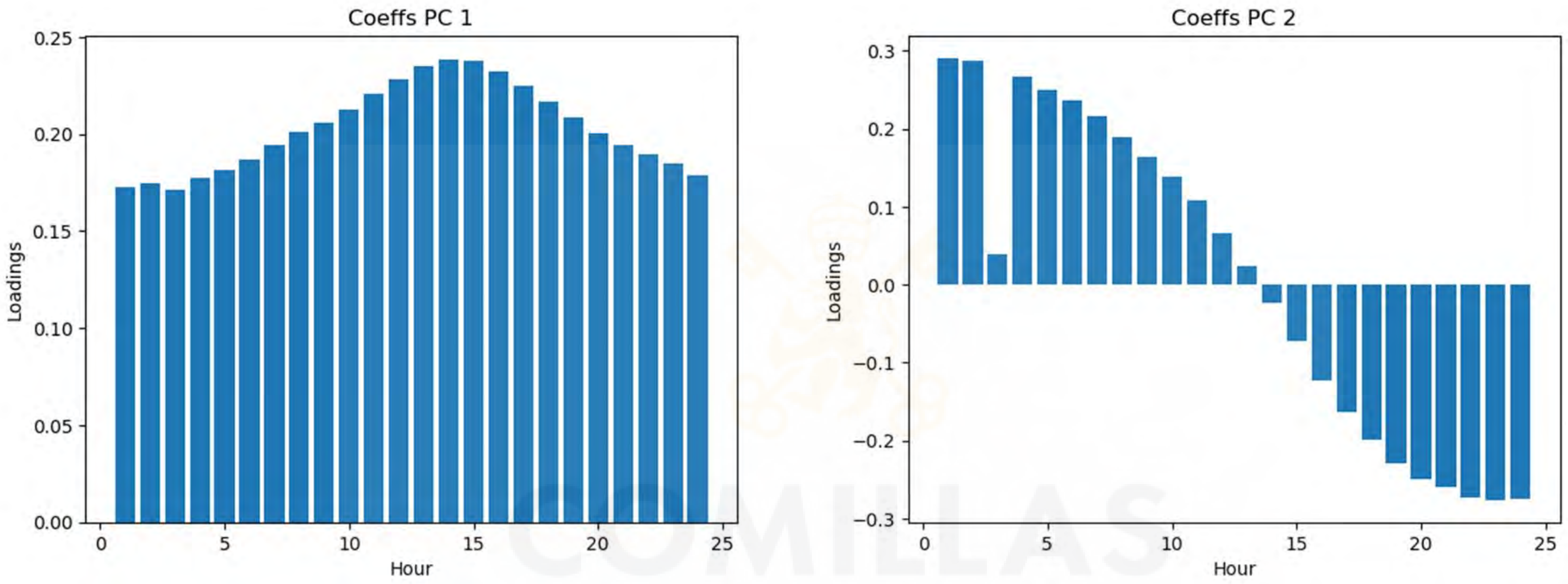
Solar PV
(Original - Reconstructed) hourly values

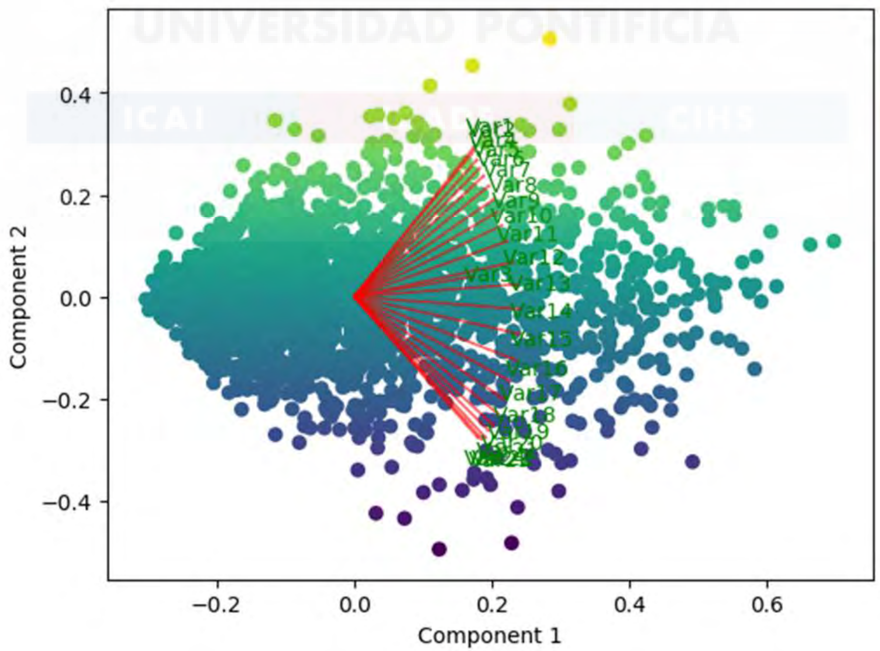


Onshore Wind
Principal Component Analysis (PCA)

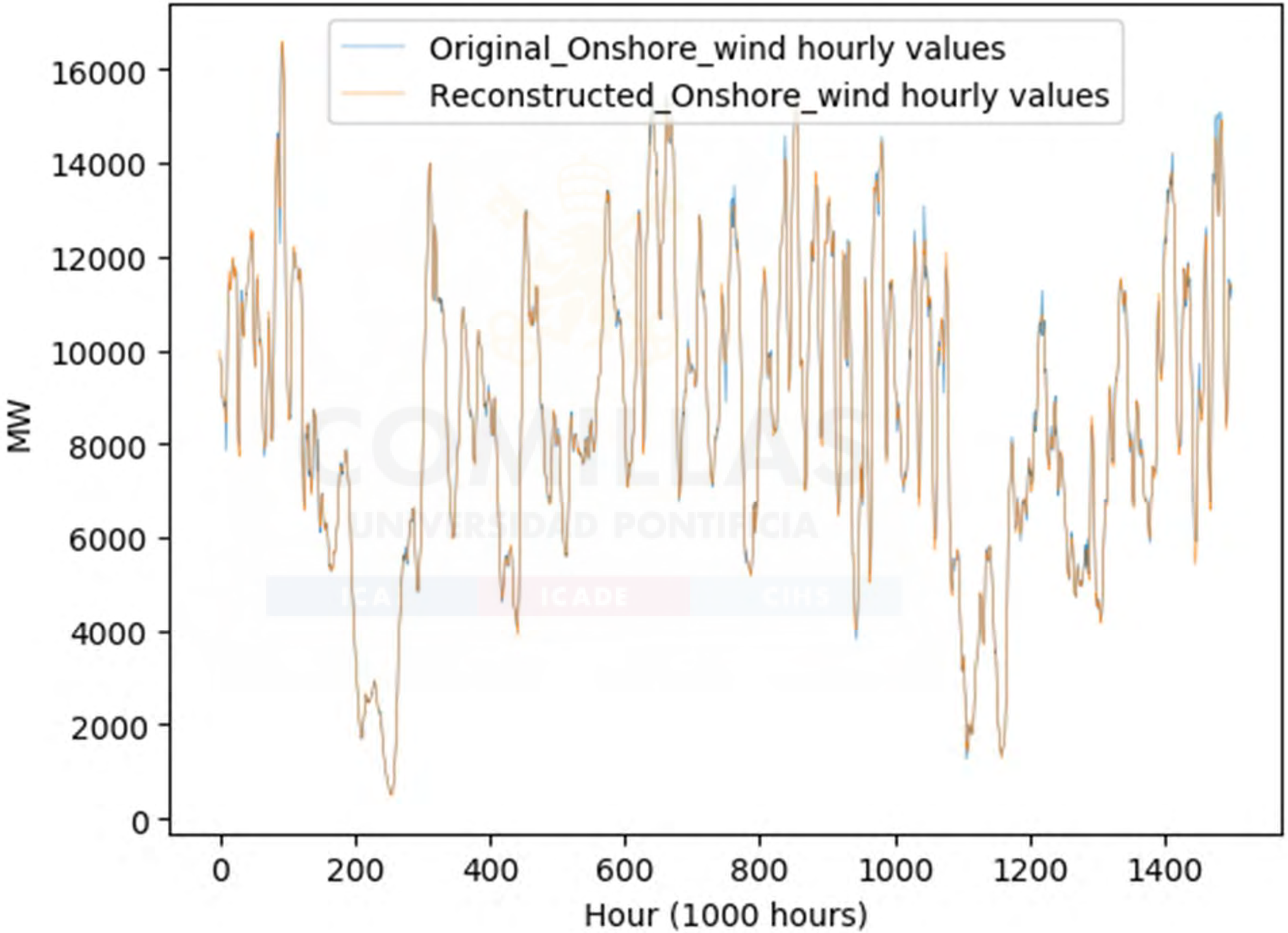


Onshore Wind
Coefficients of PC1 and PC2 and scores

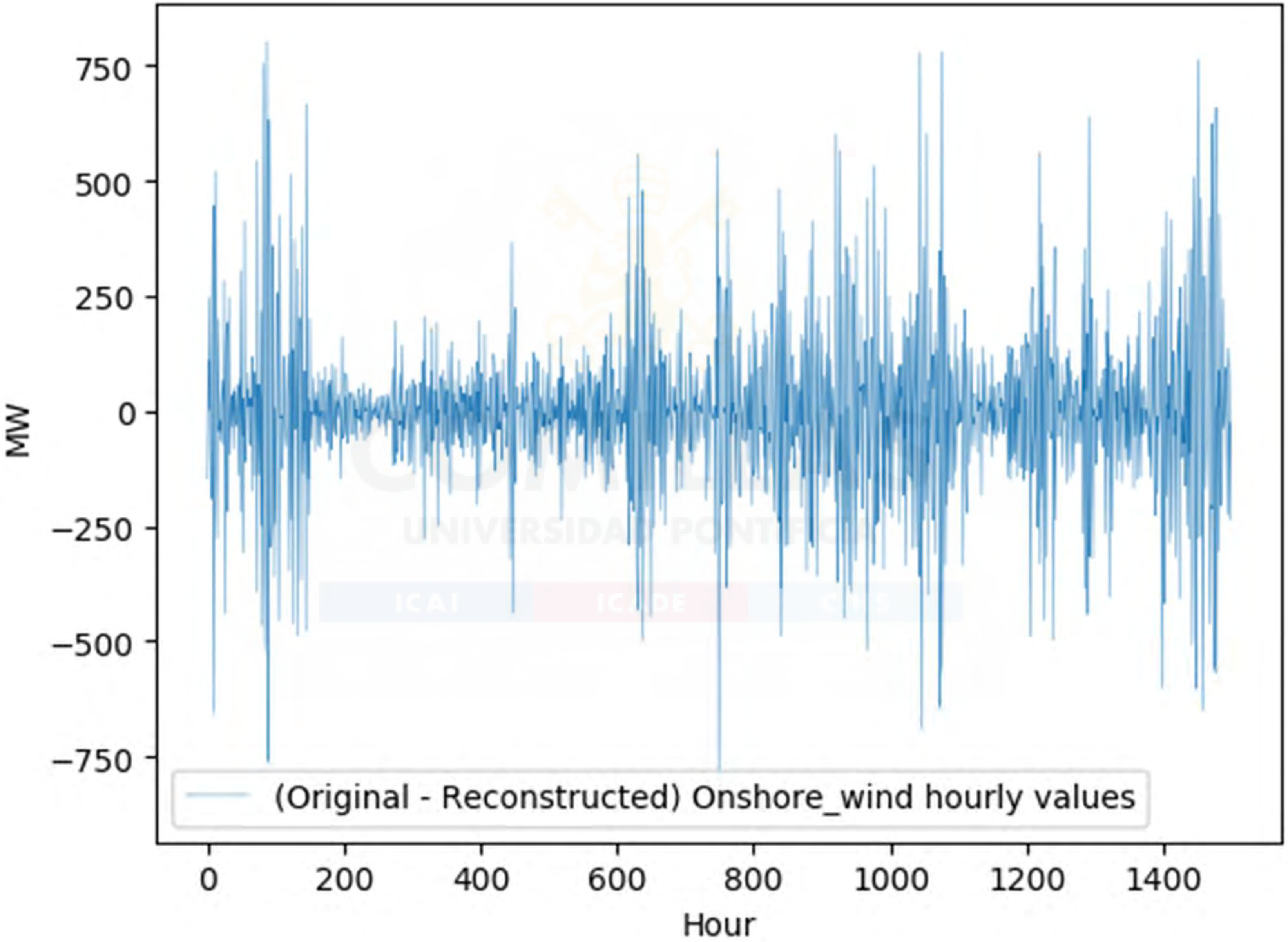




Onshore Wind
Original and reconstructed hourly values with 2 PCs



Onshore Wind
(Original - Reconstructed) hourly values



5

Representative patterns

Find daily generation patterns for each technology

- 24 values describe hourly output
- We want to find typical representative days of the 5-year daily output (2191 days)

Clustering
Introduction

- Clustering refers to a comprehensive set of techniques for finding subgroups, or clusters, in a data set

Each observation is a vehicle described by input variables (features) such as the weight, mean speed, and number of wheels.

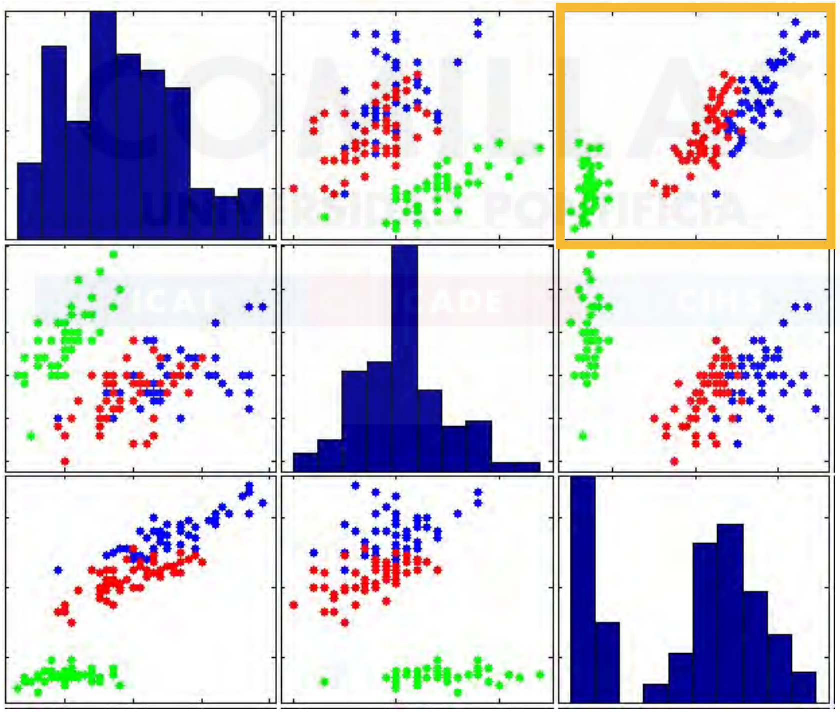
When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.

What does it mean to be *similar* or *different?*

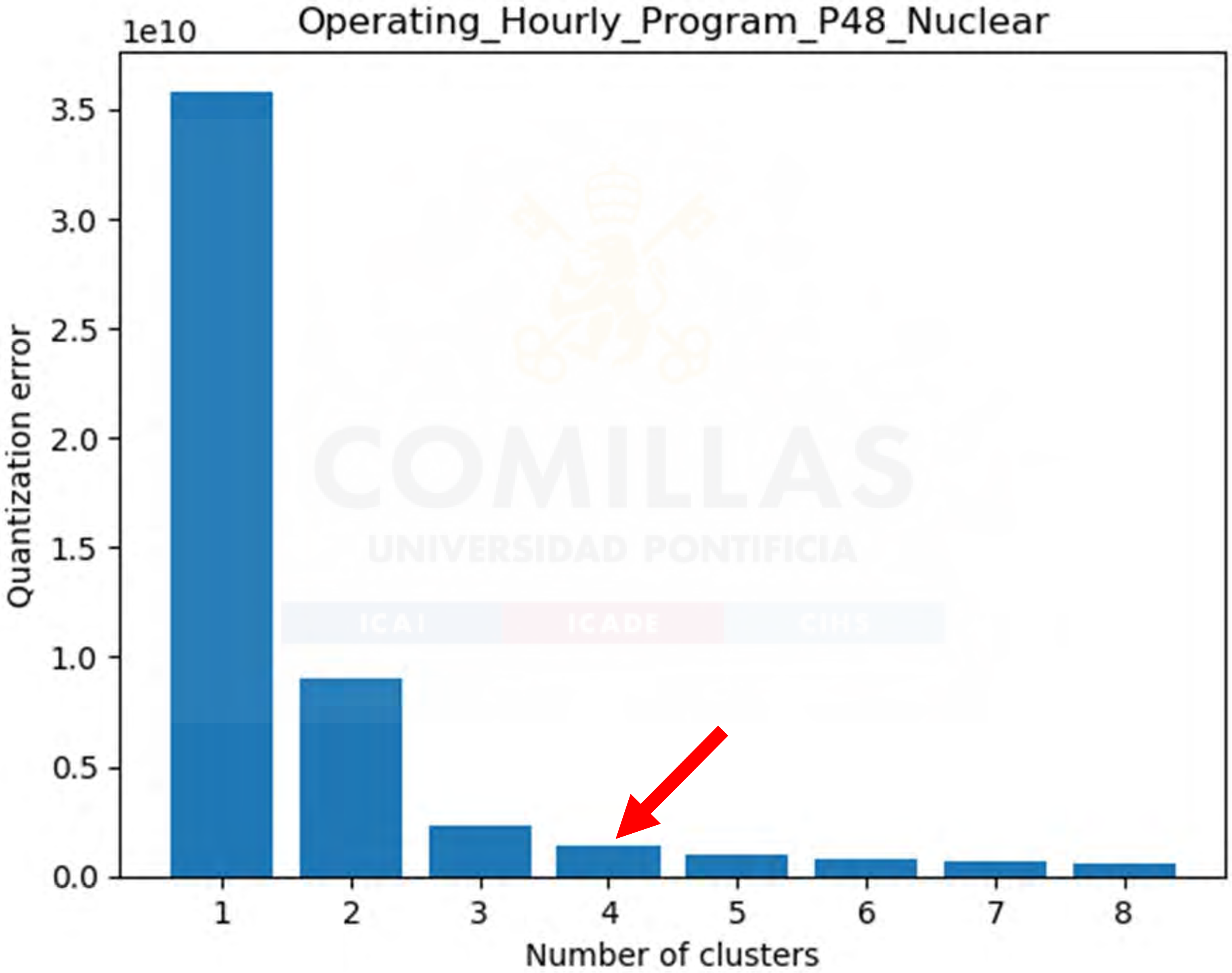
Unlabeled observations (no output variable)

Clustering
Introduction

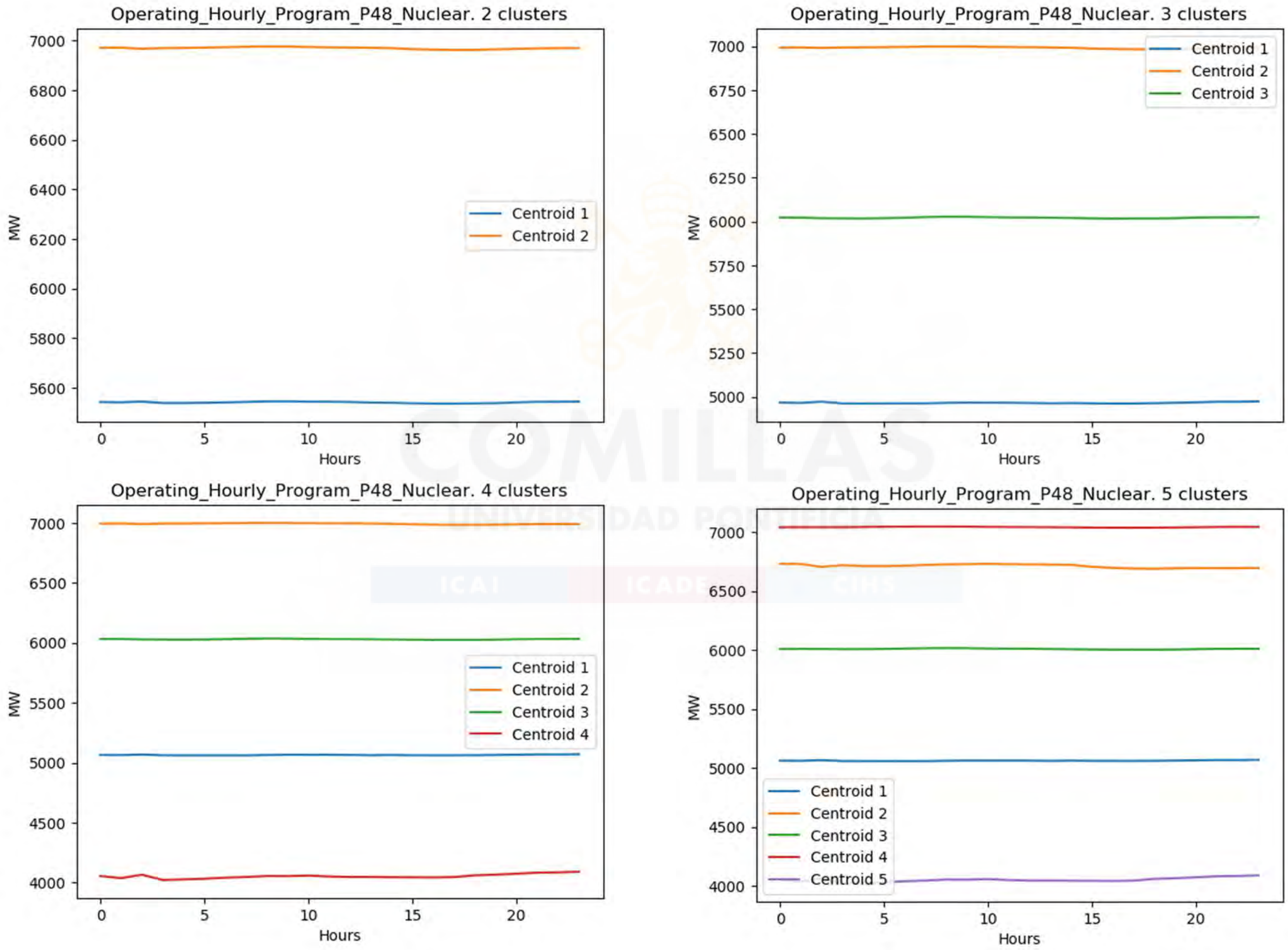
- Why clustering?
  - To gain some insight into the structure of the data (interpretation)
  - Discovery of patterns
  - Grouping highly correlated attributes



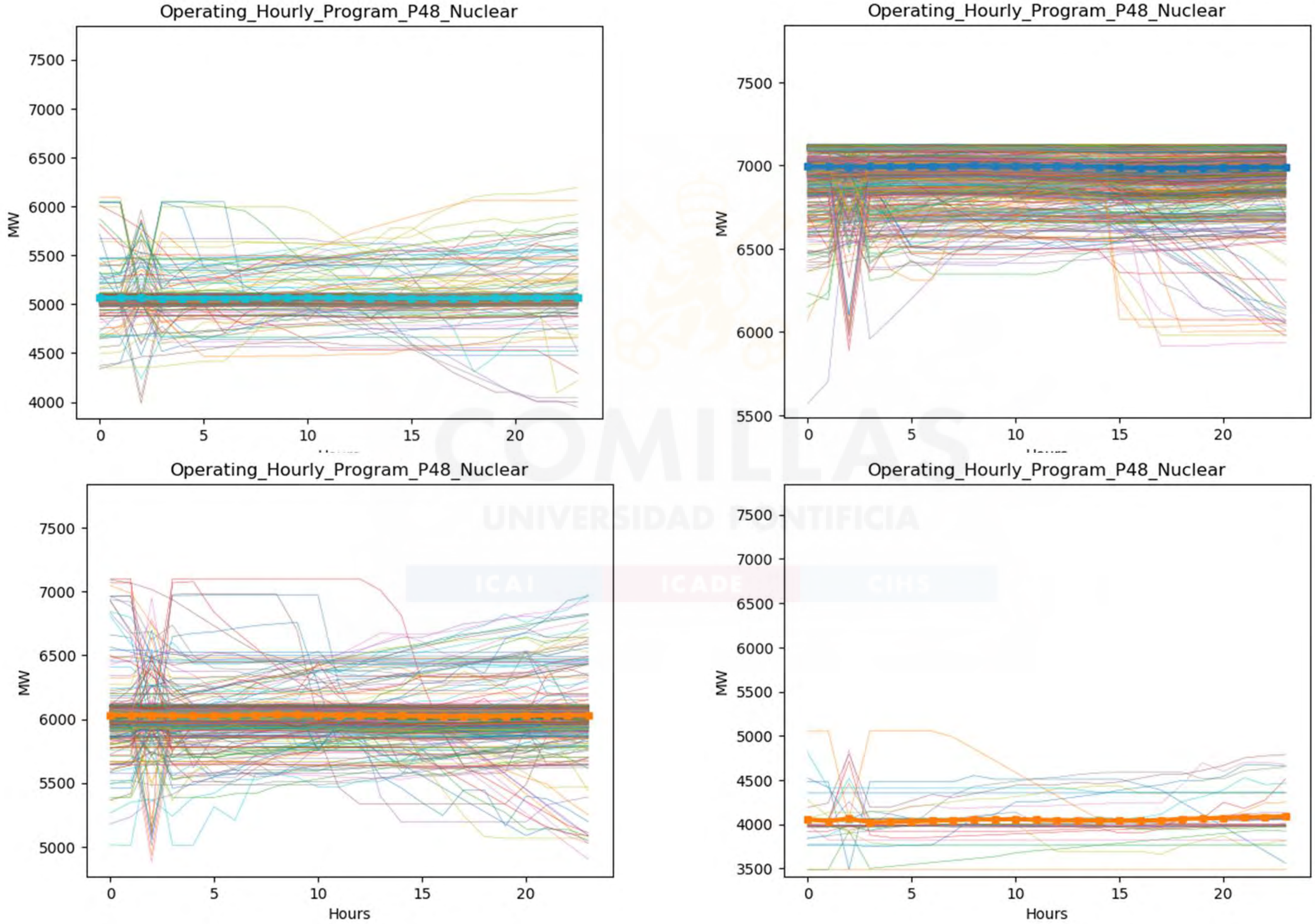
Hourly Nuclear
Quantization error. Intracluster distance



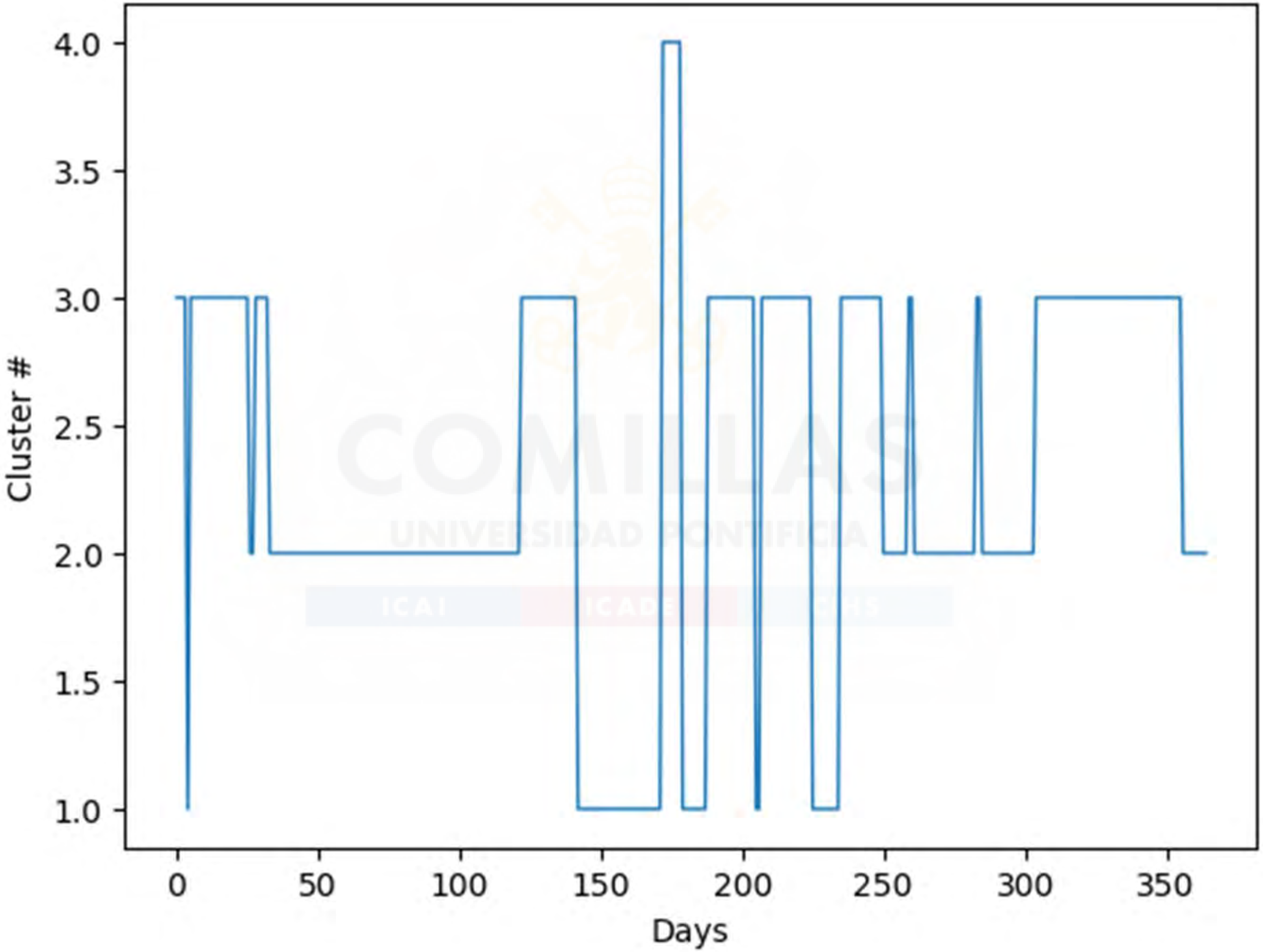
Hourly Nuclear
2, 3, 4 and 5 clusters



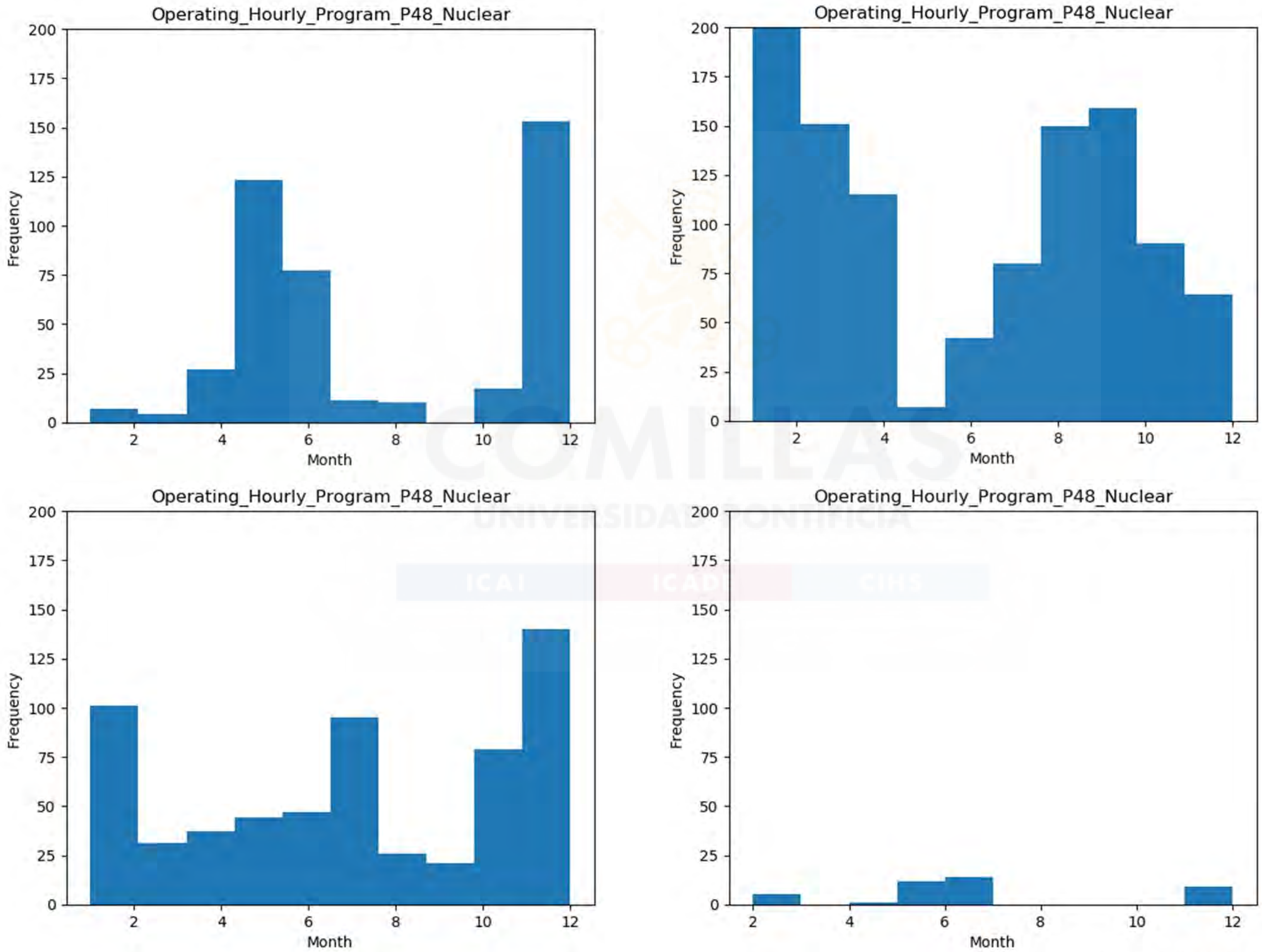
Hourly Nuclear
Clusters 1, 2, 3, and 4



Hourly Nuclear
Cluster order (for the first year)

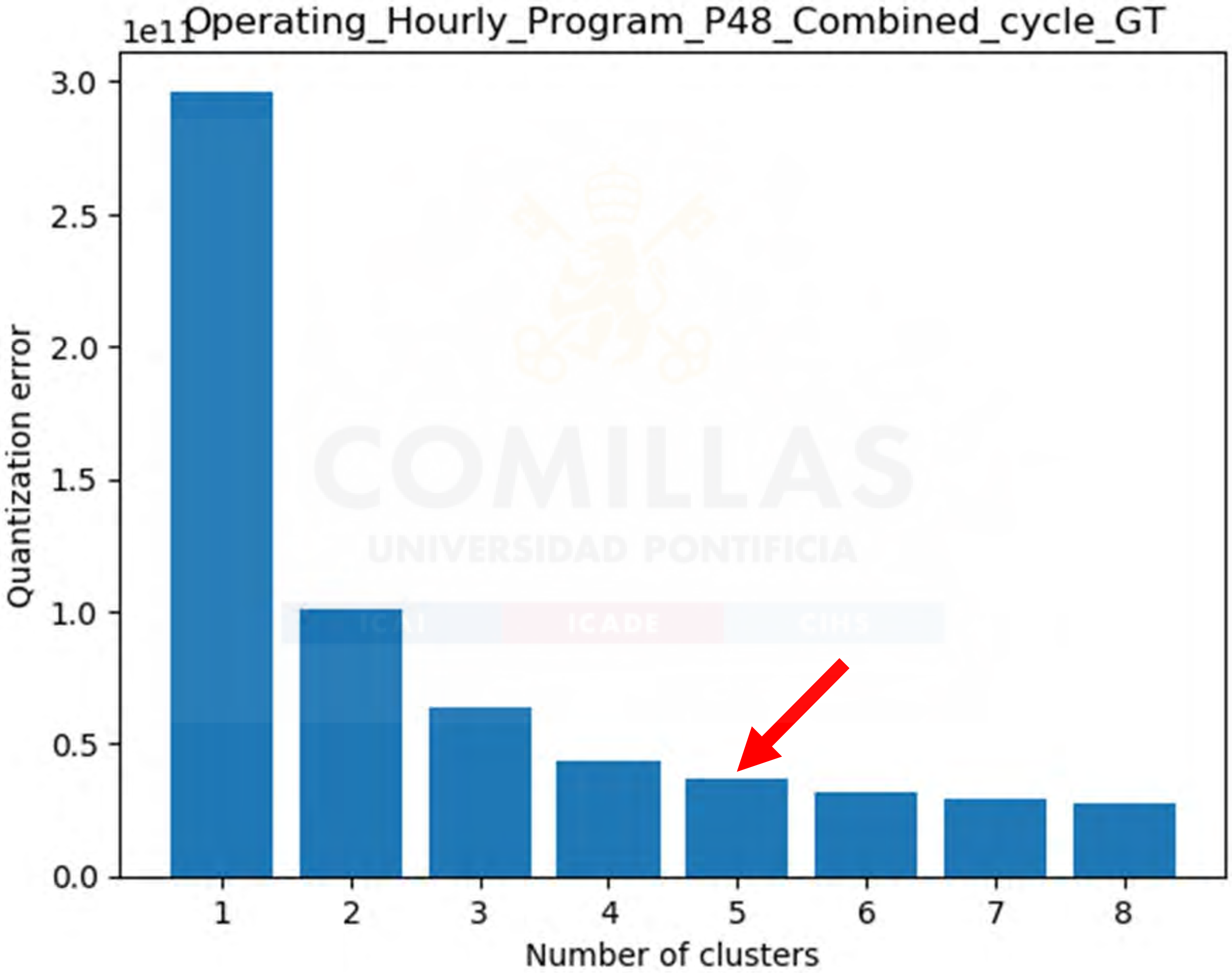


Hourly Nuclear
Clusters 1, 2, 3, and 4



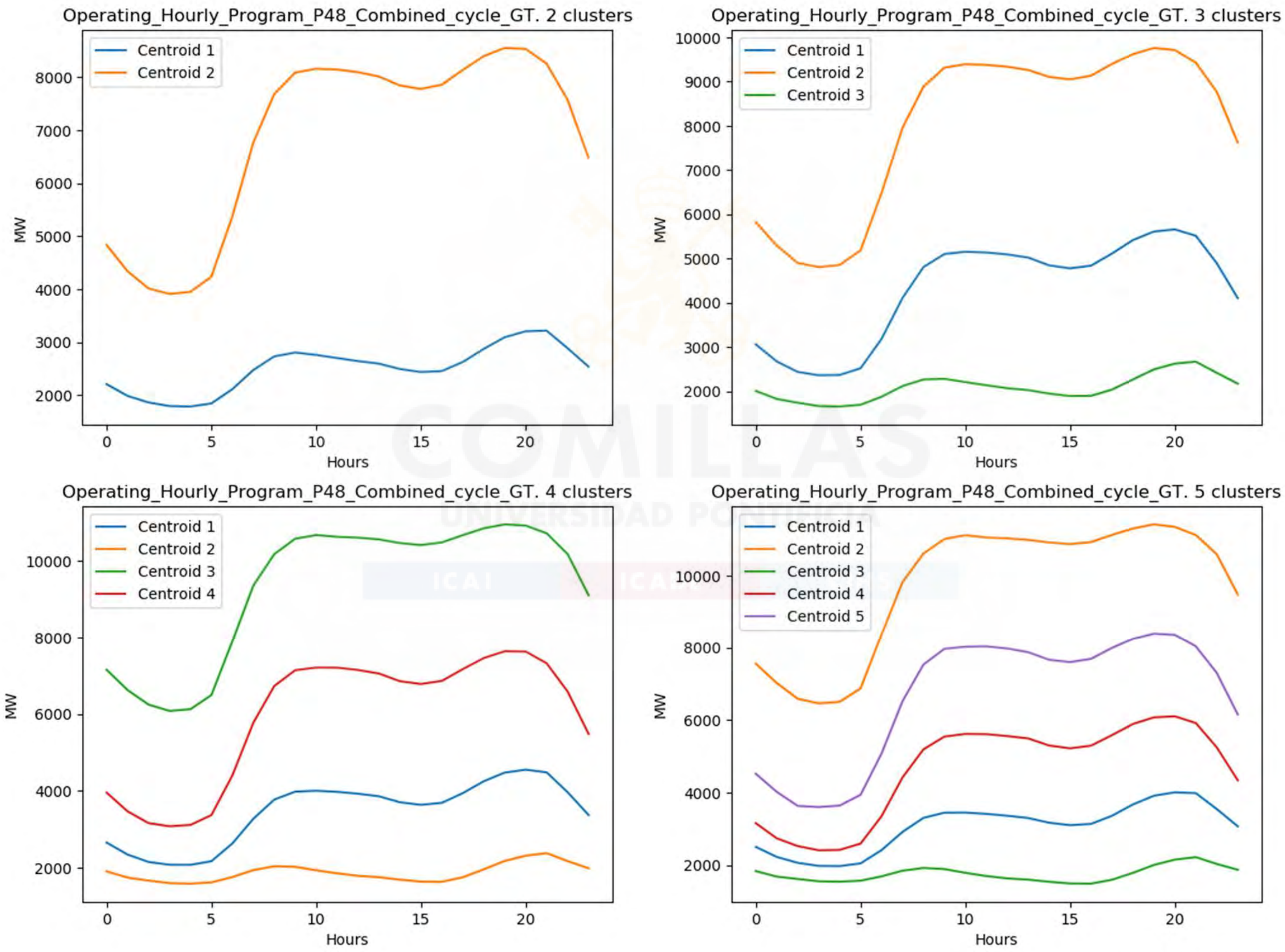
Hourly Combined Cycle GT

Quantization error. Intracluster distance

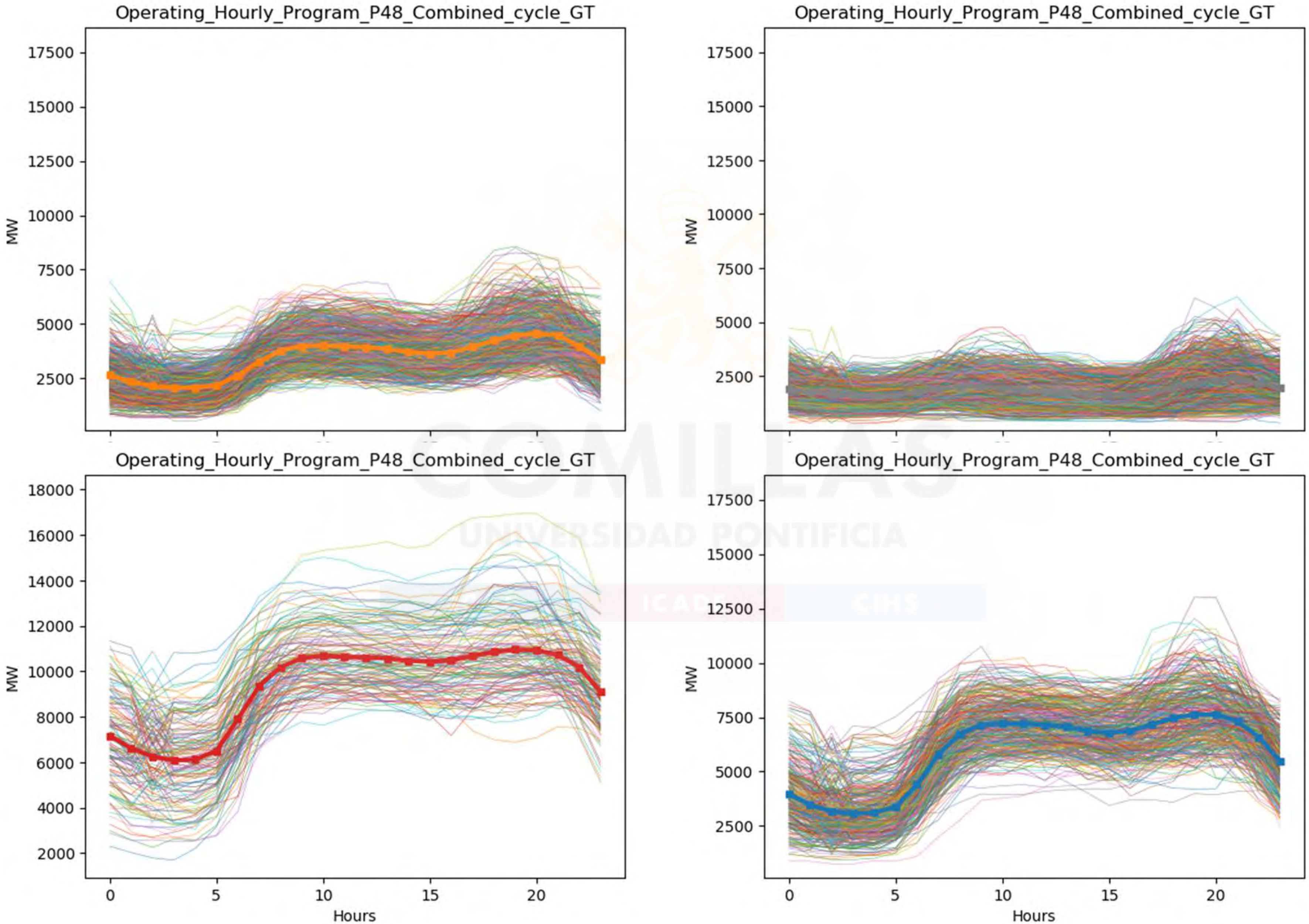


Hourly Combined Cycle GT

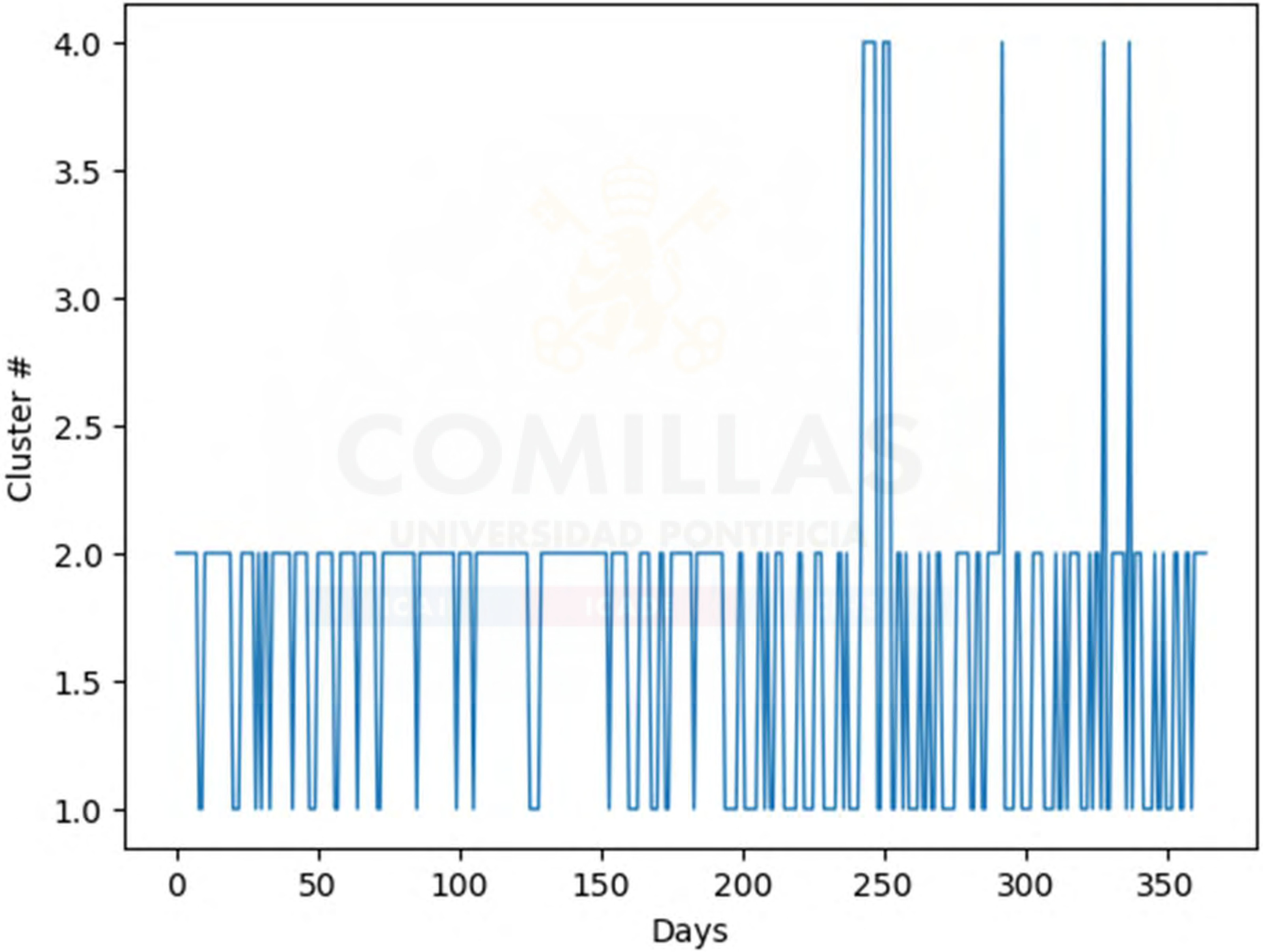
2, 3, 4 and 5 clusters



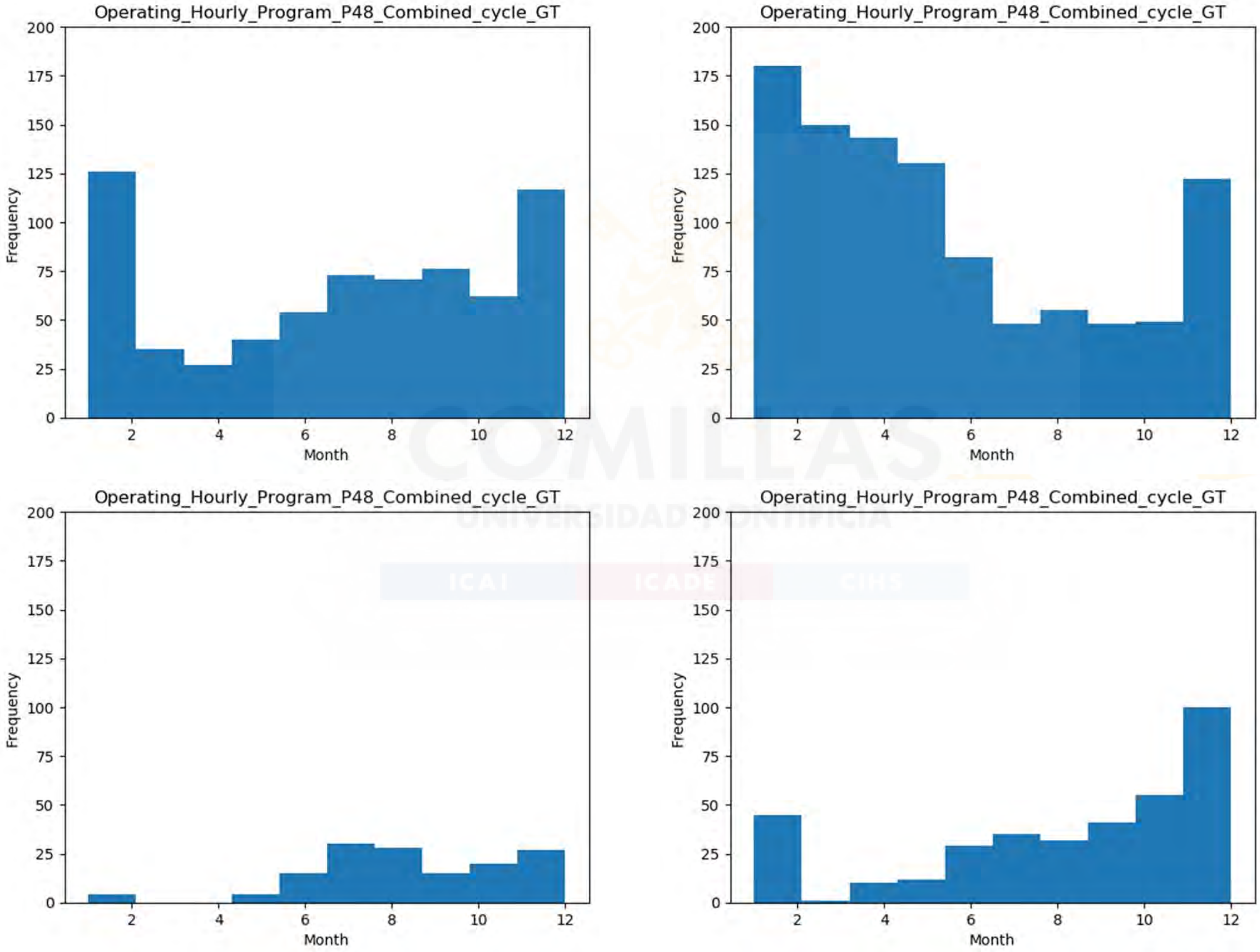
Hourly Combined Cycle GT
Clusters 1, 2, 3, and 4



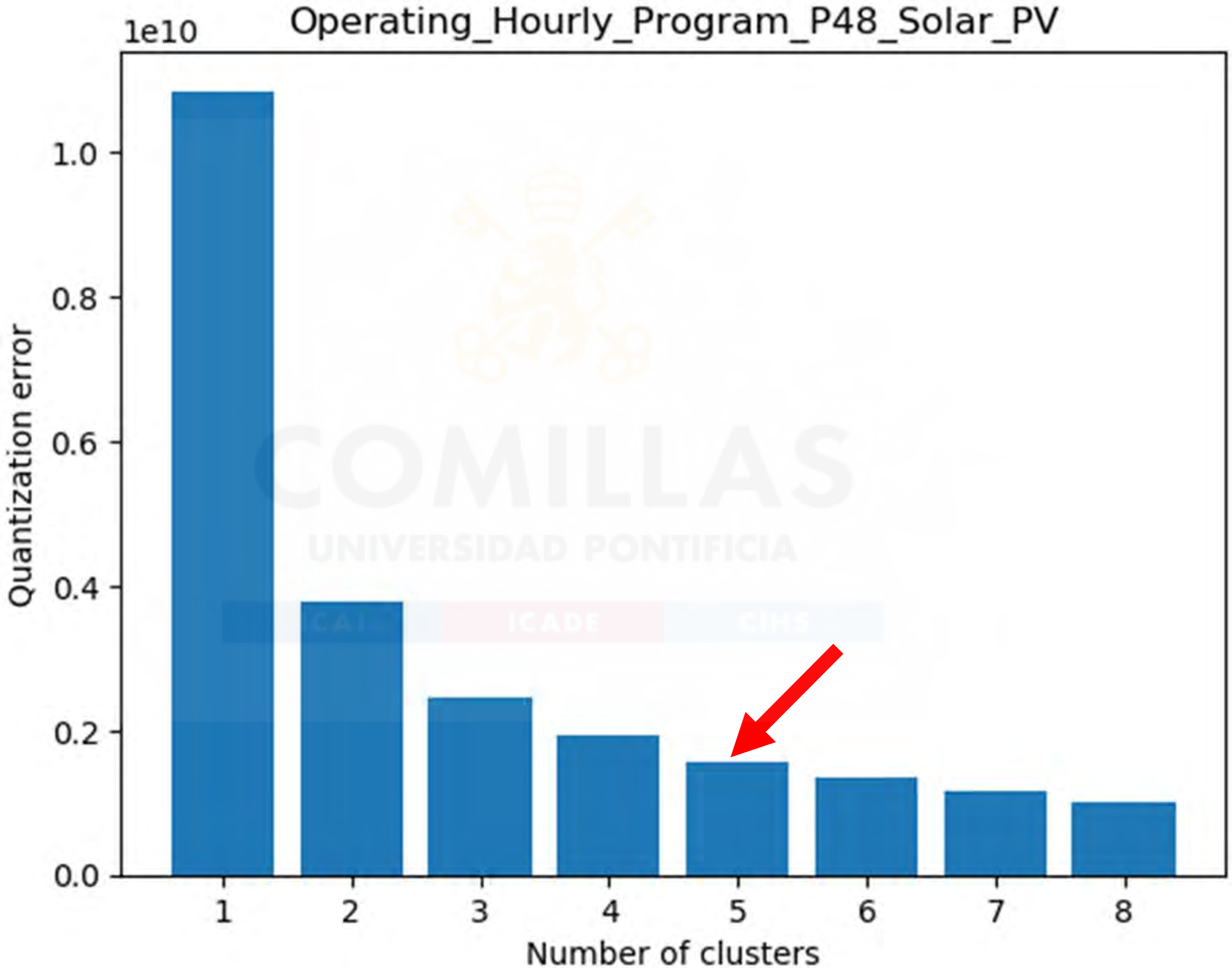
Hourly Combined Cycle GT
Cluster order (for the first year)



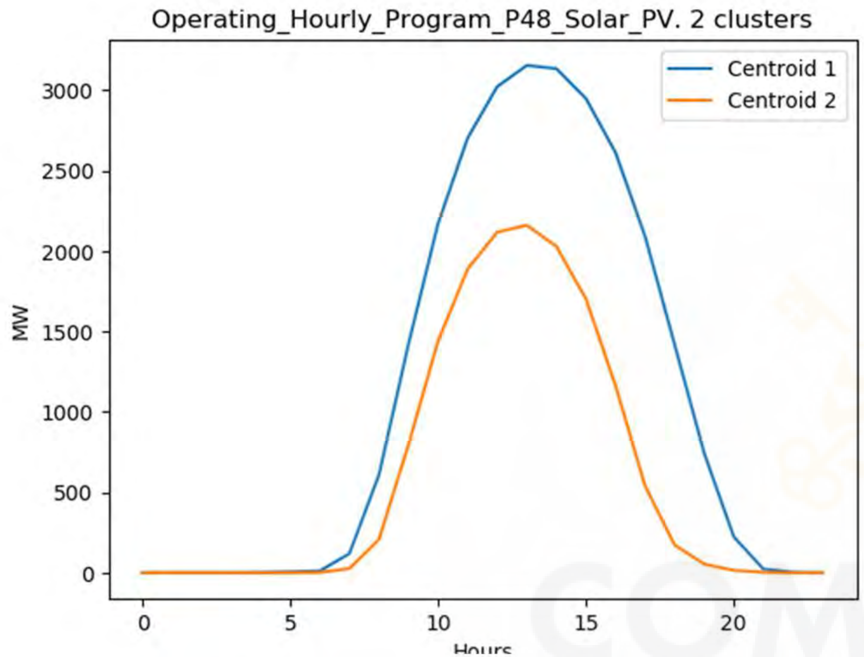
Hourly Combined Cycle GT
Clusters 1, 2, 3, and 4

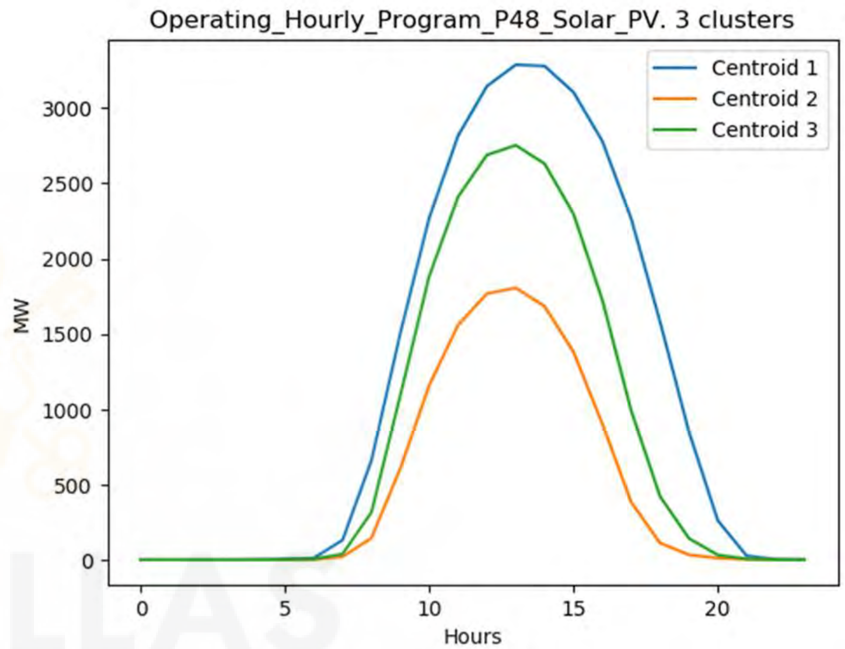


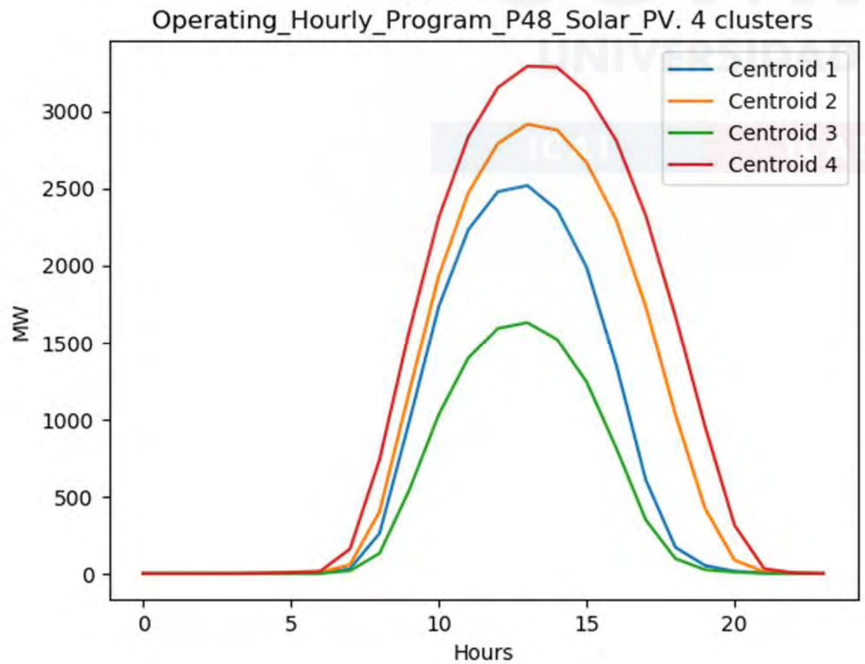
Hourly Solar PV
Quantization error. Intracluster distance

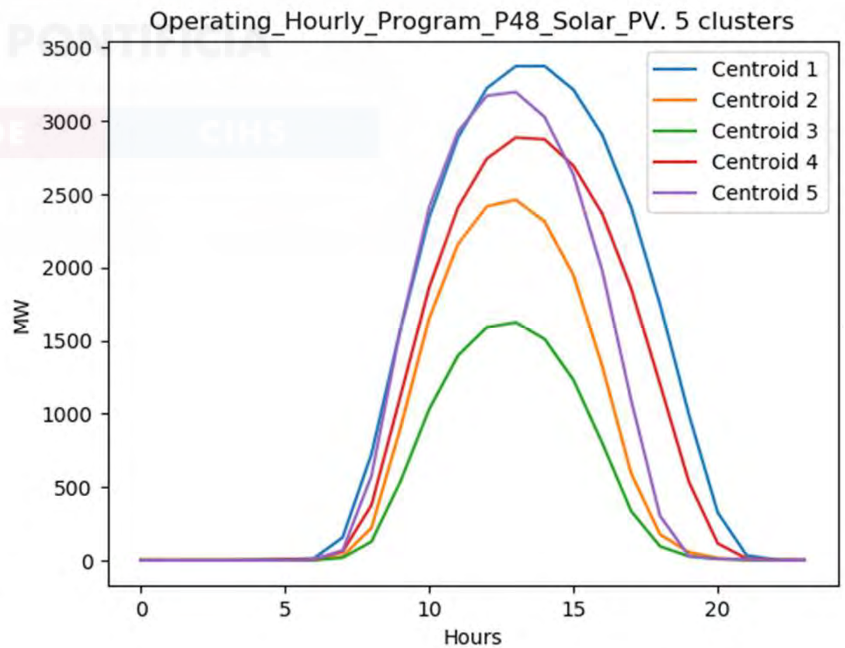


Hourly Solar PV
2, 3, 4 and 5 clusters

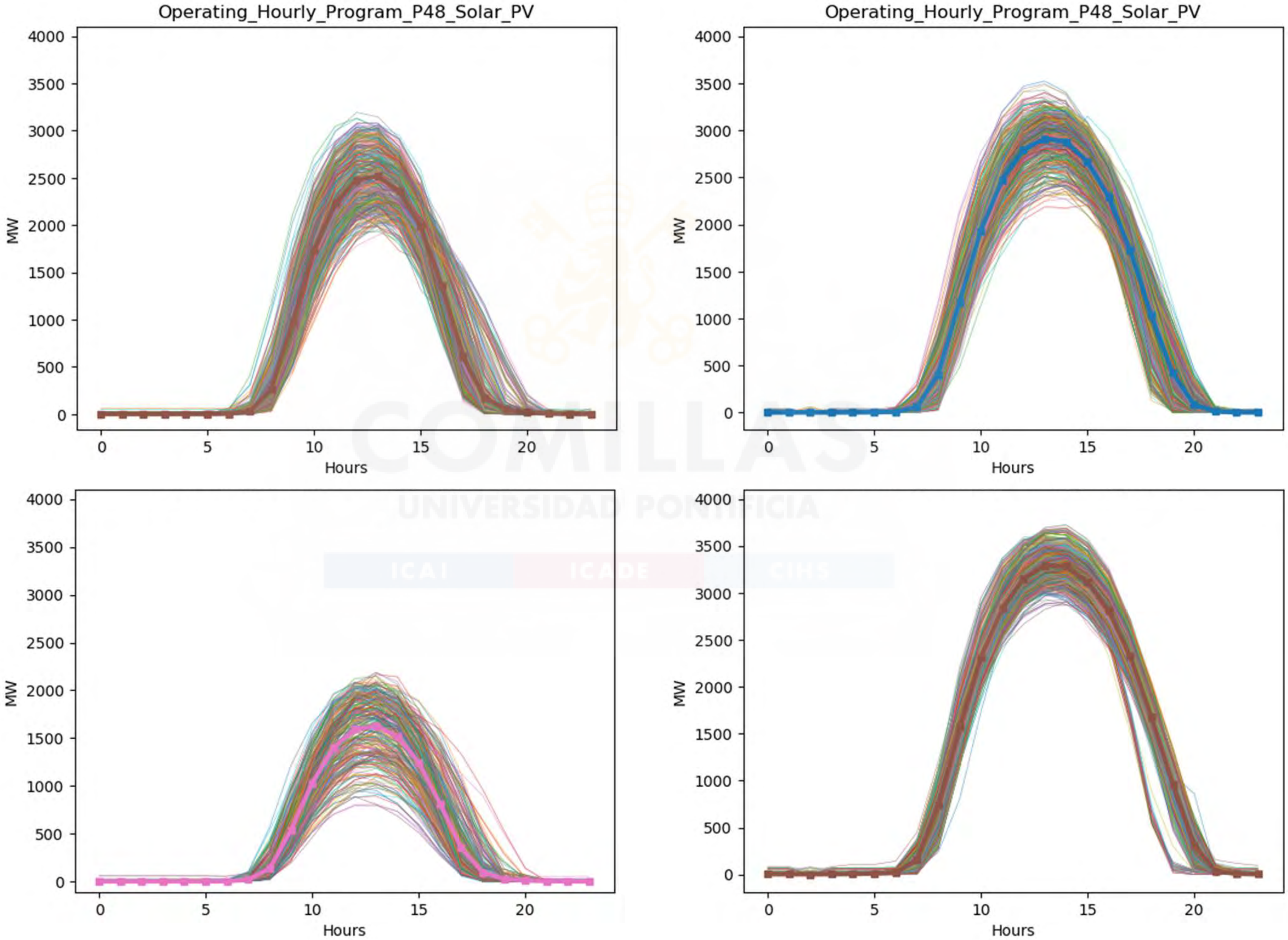






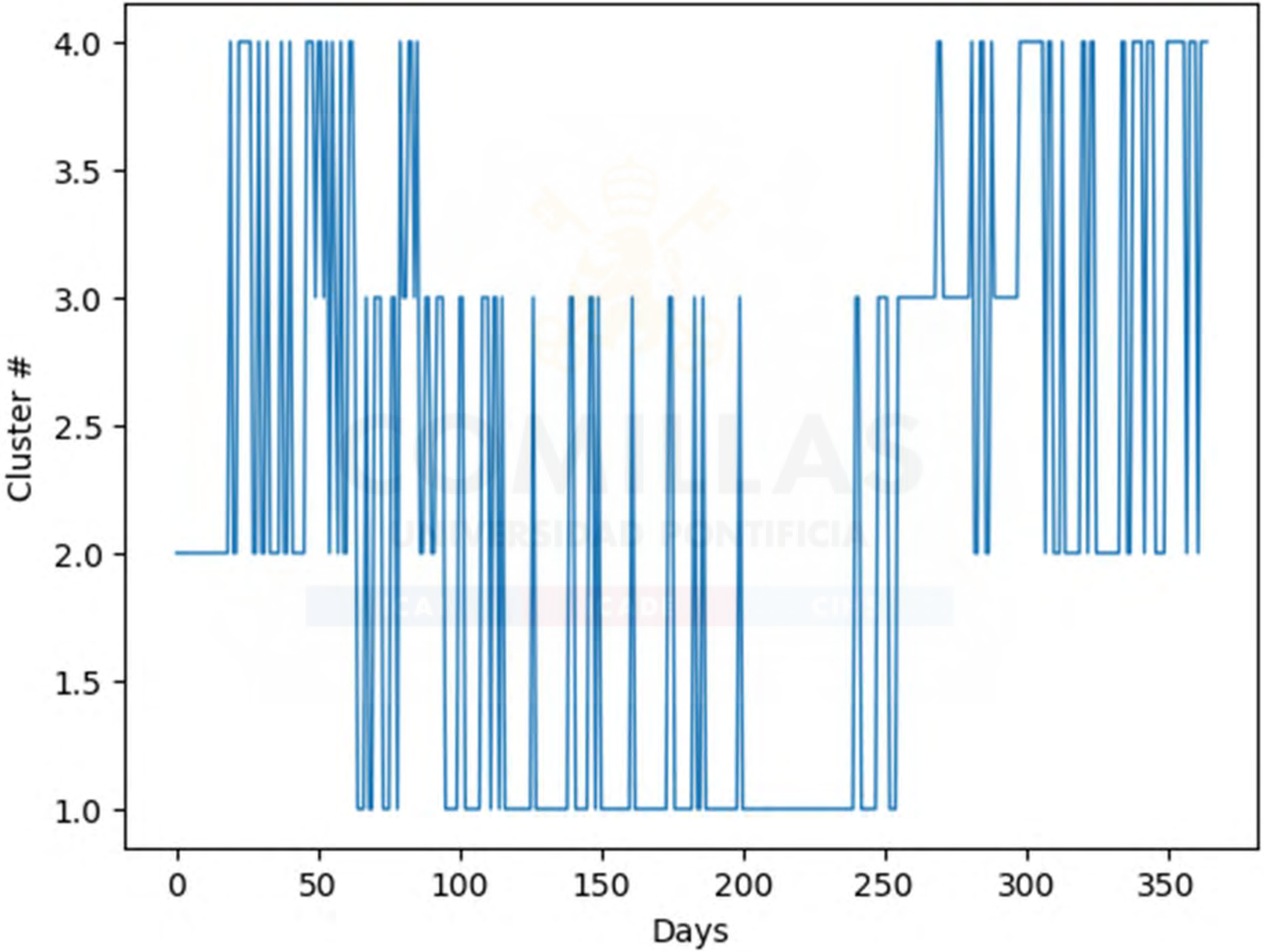


Hourly Solar PV
Clusters 1, 2, 3, and 4

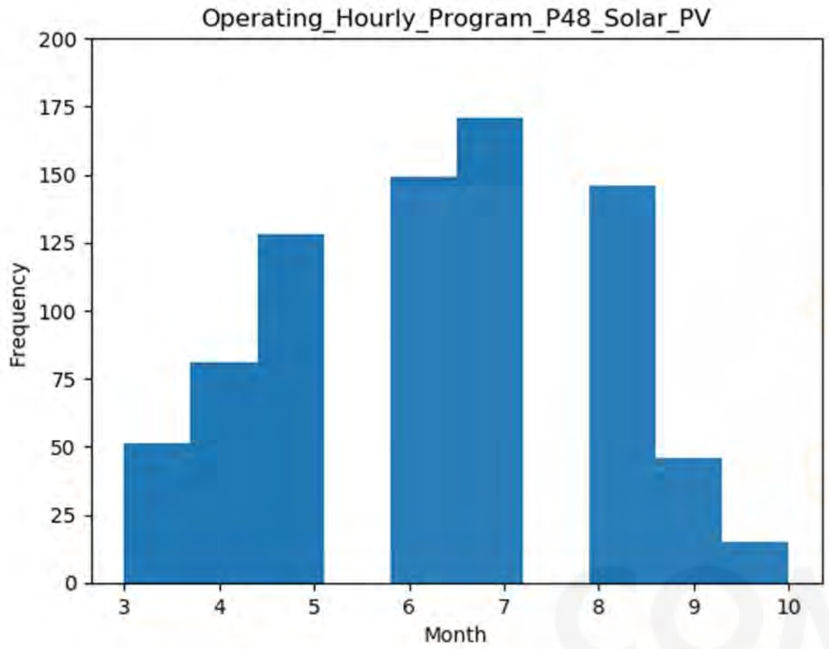


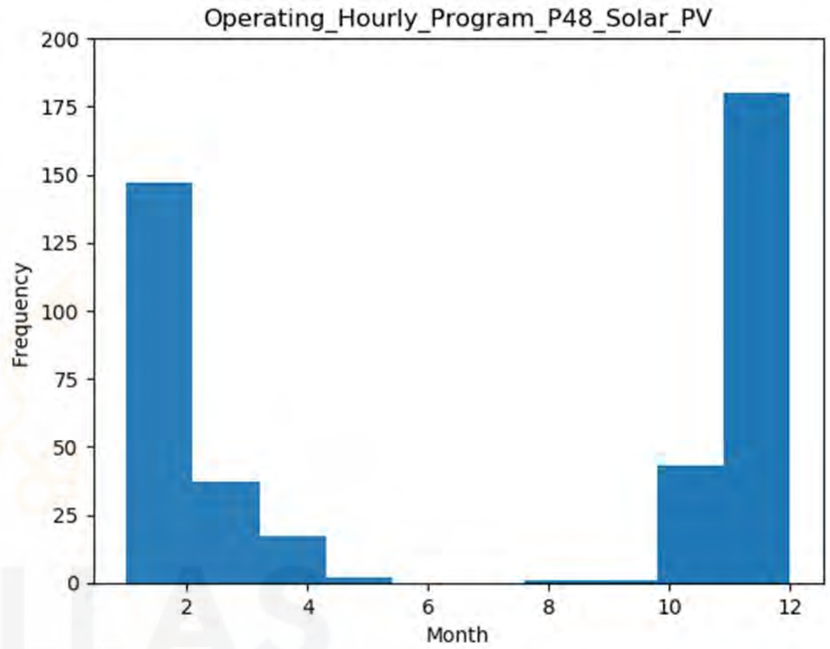
Hourly Solar PV

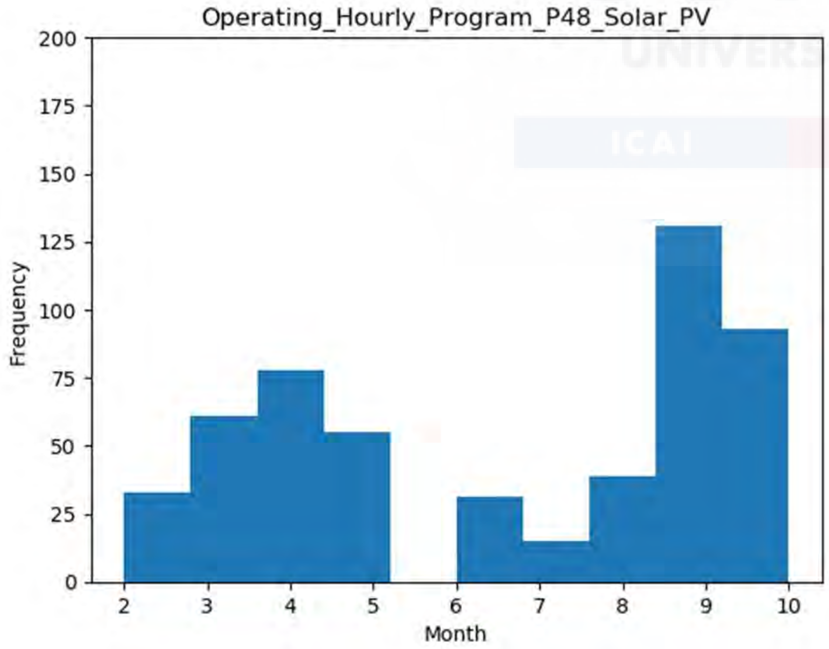
Cluster order (for the first year)

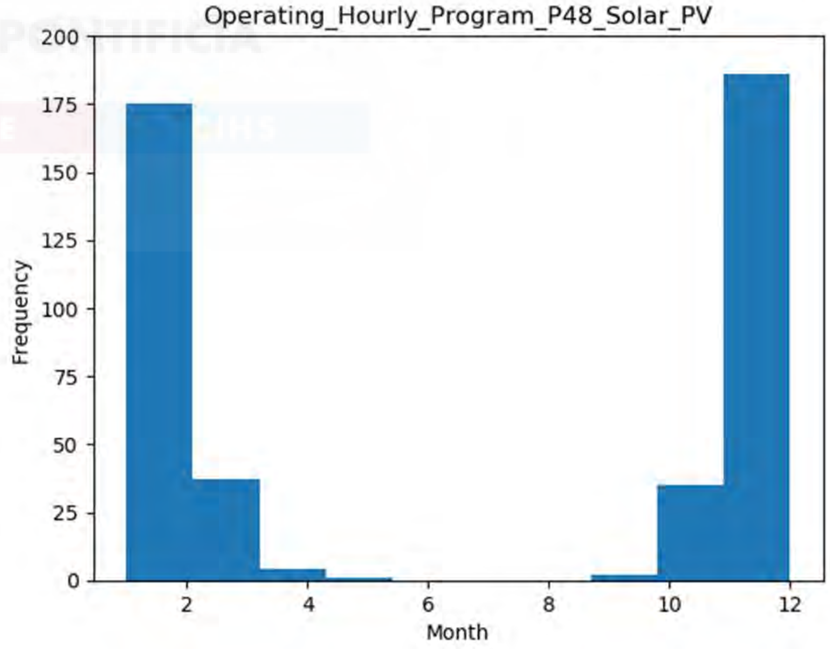


Hourly Solar PV
Clusters 1, 2, 3, and 4

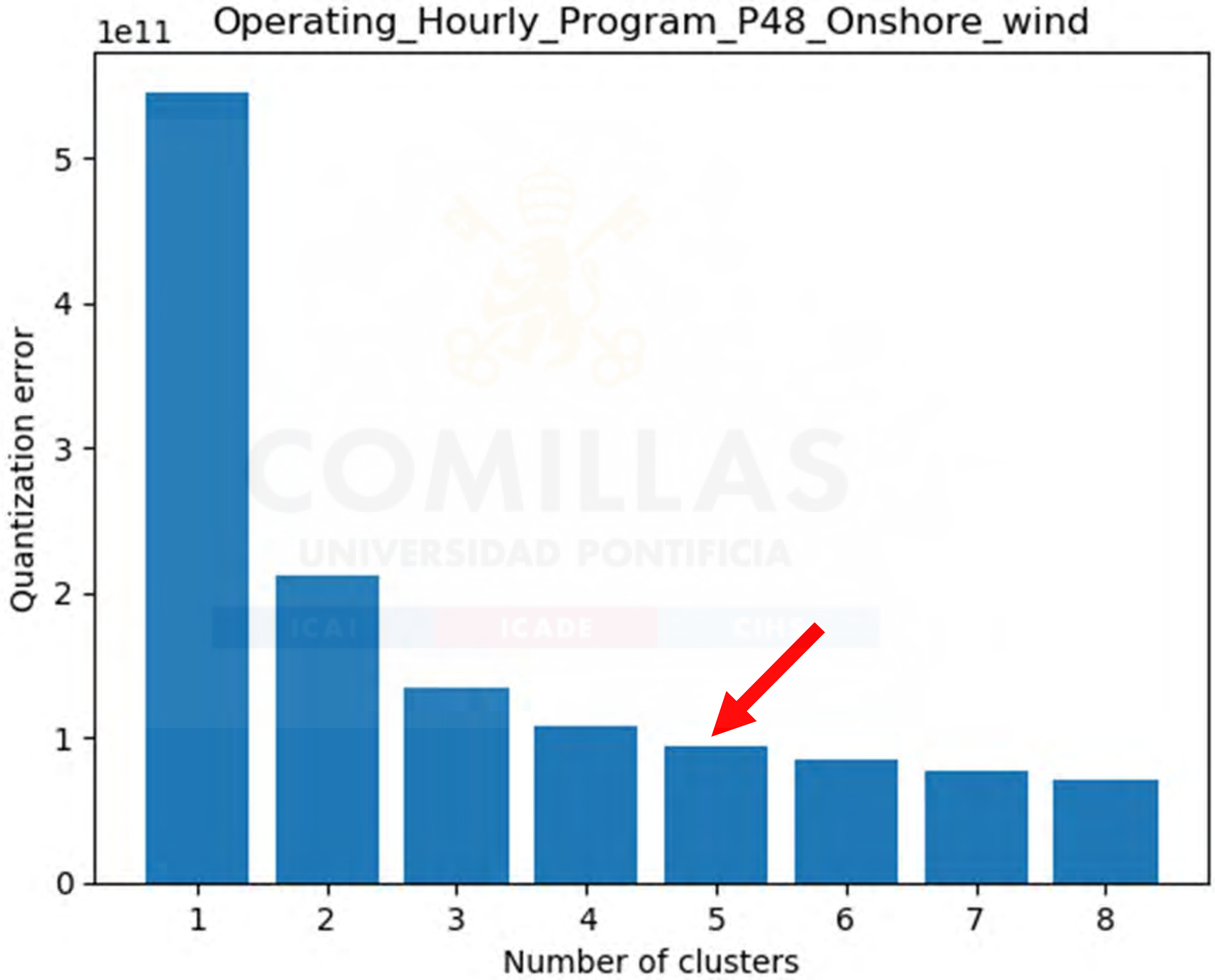




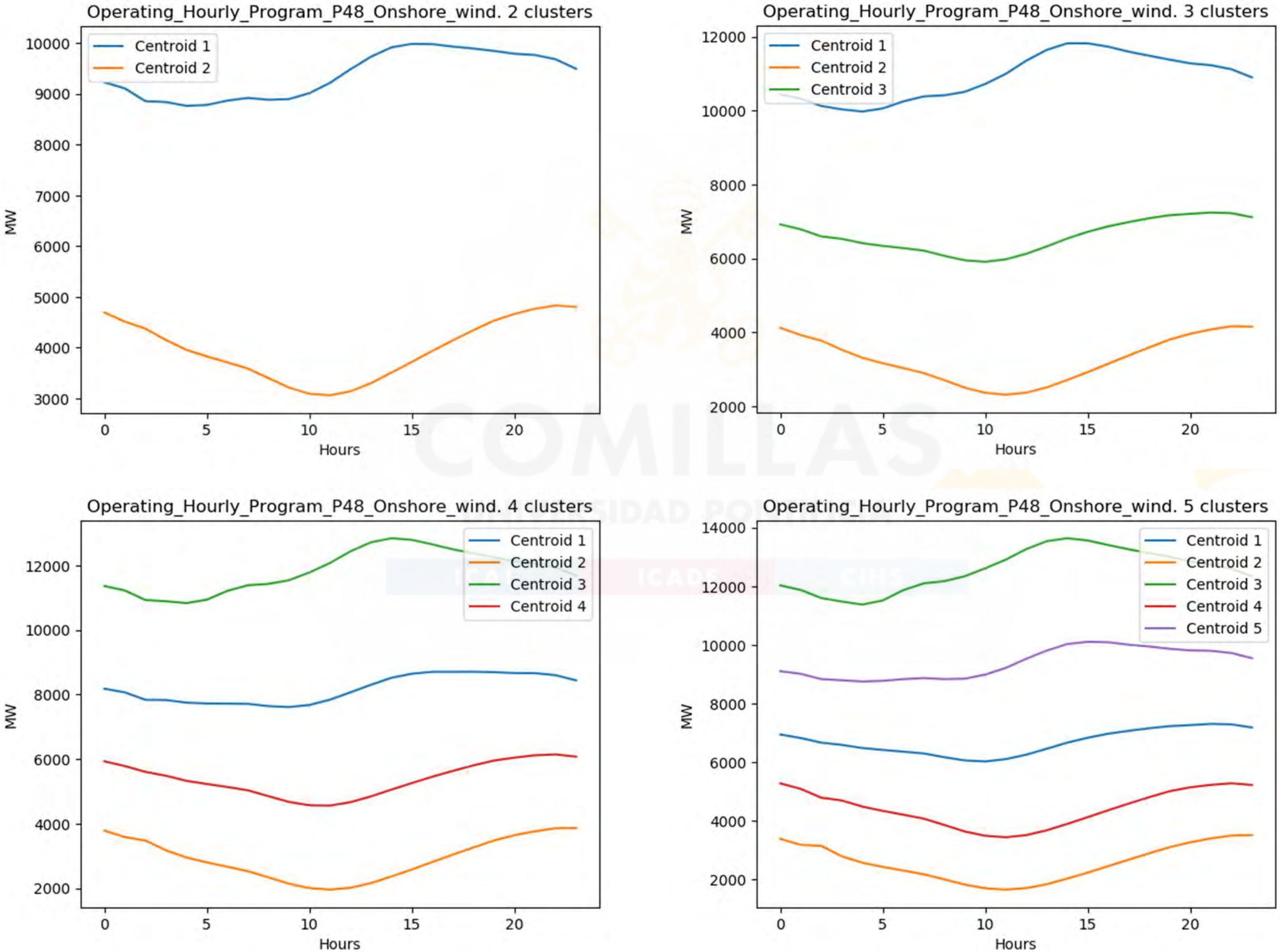




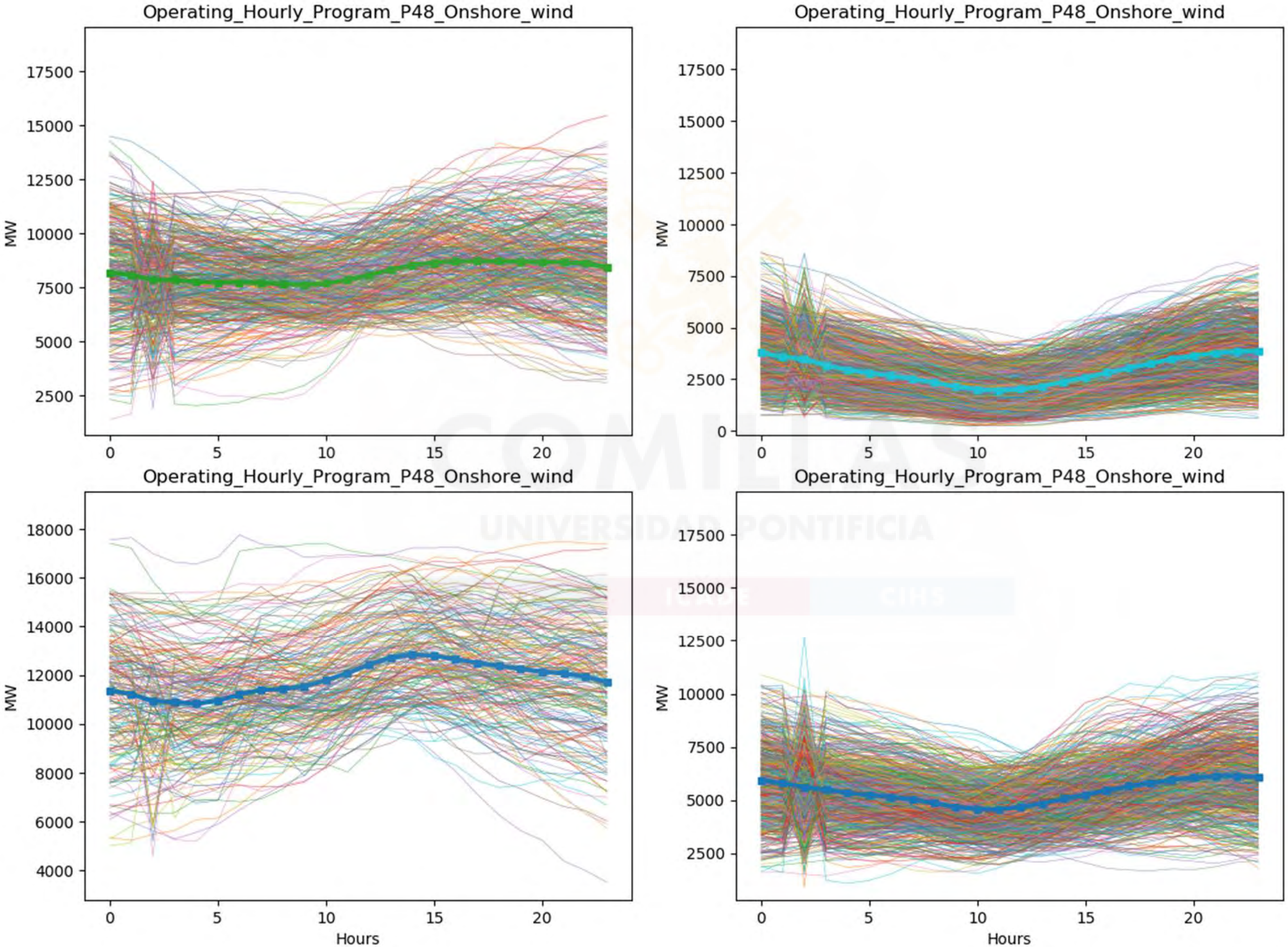
Hourly Onshore Wind
Quantization error. Intracluster distance



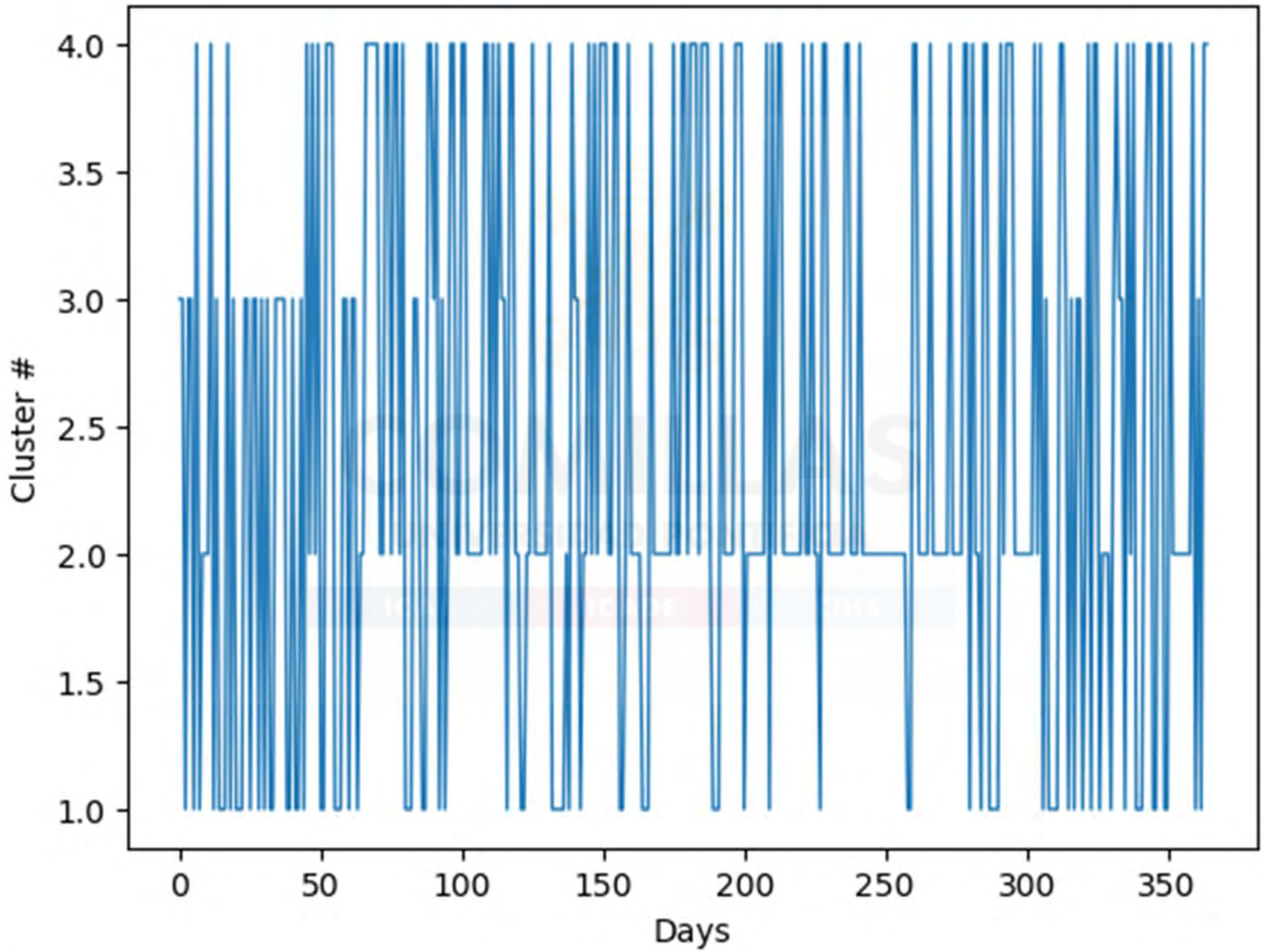
Hourly Onshore Wind
2, 3, 4 and 5 clusters



Hourly Onshore Wind
Clusters 1, 2, 3, and 4

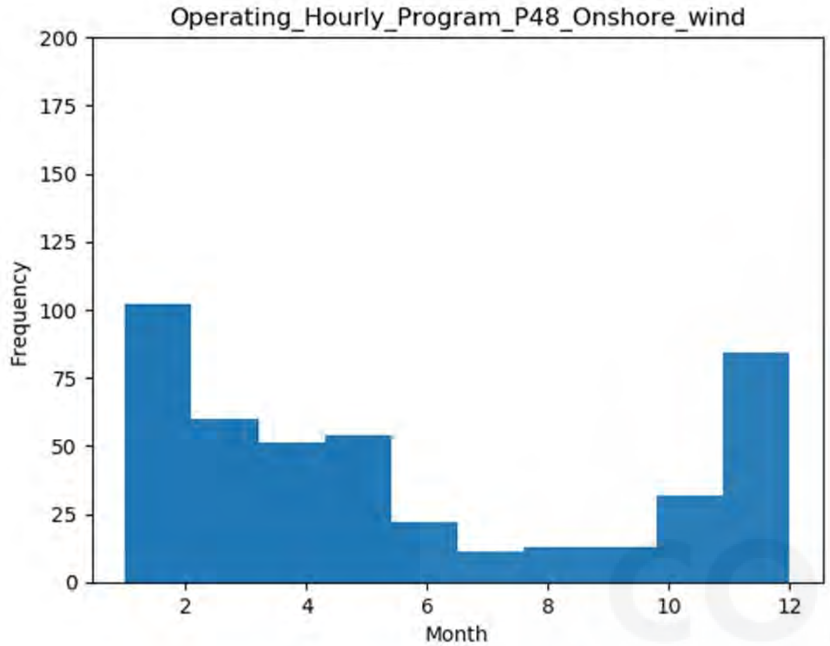


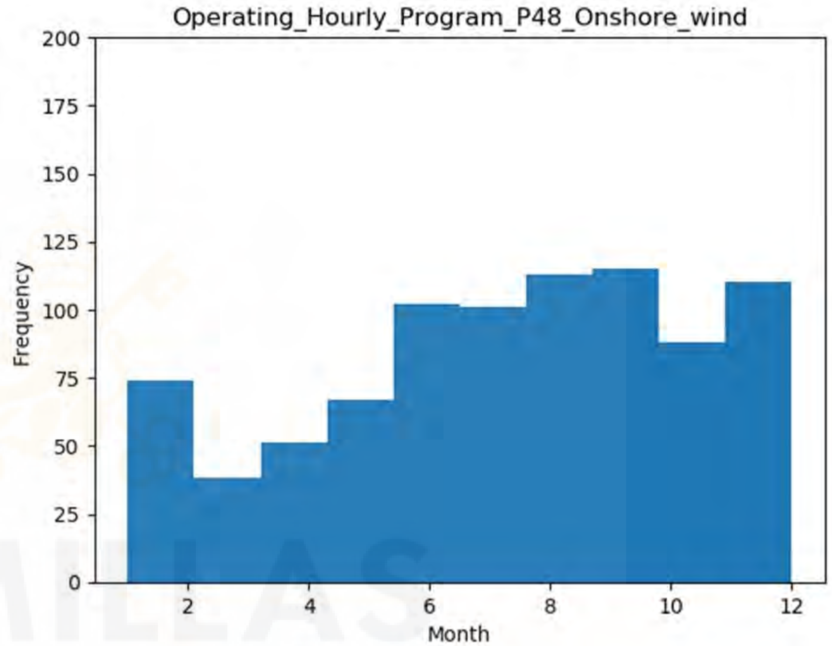
Hourly Onshore Wind
Cluster order (for the first year)

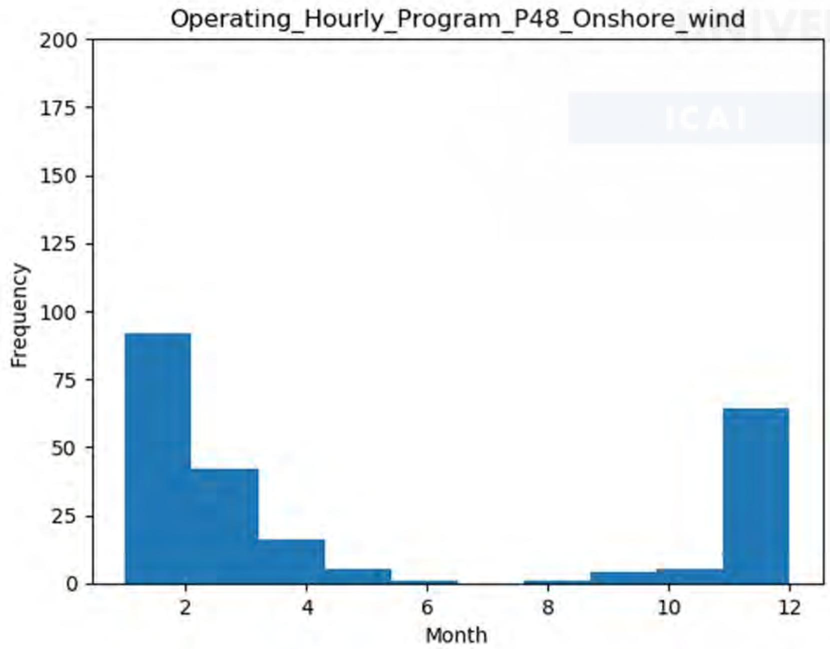


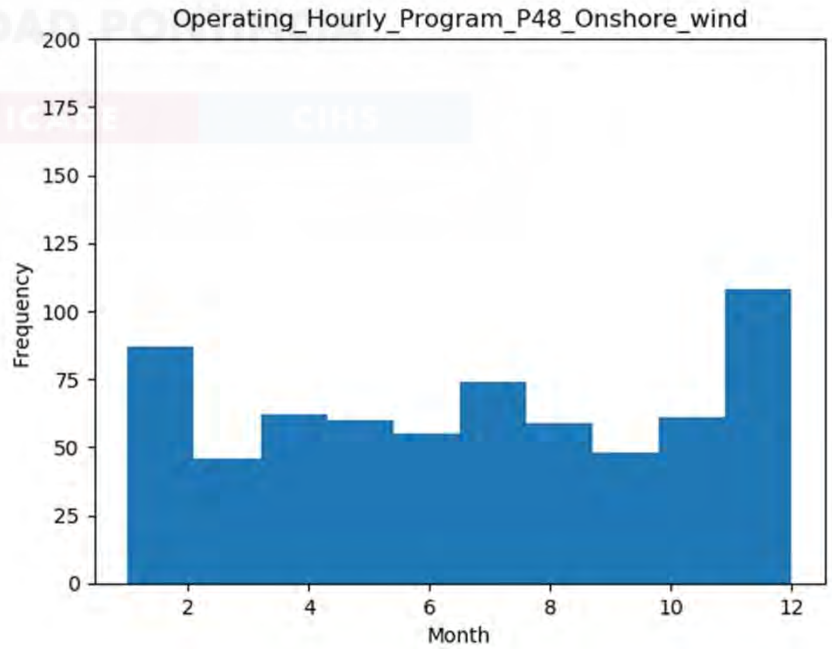
Hourly Onshore Wind

Clusters 1, 2, 3, and 4









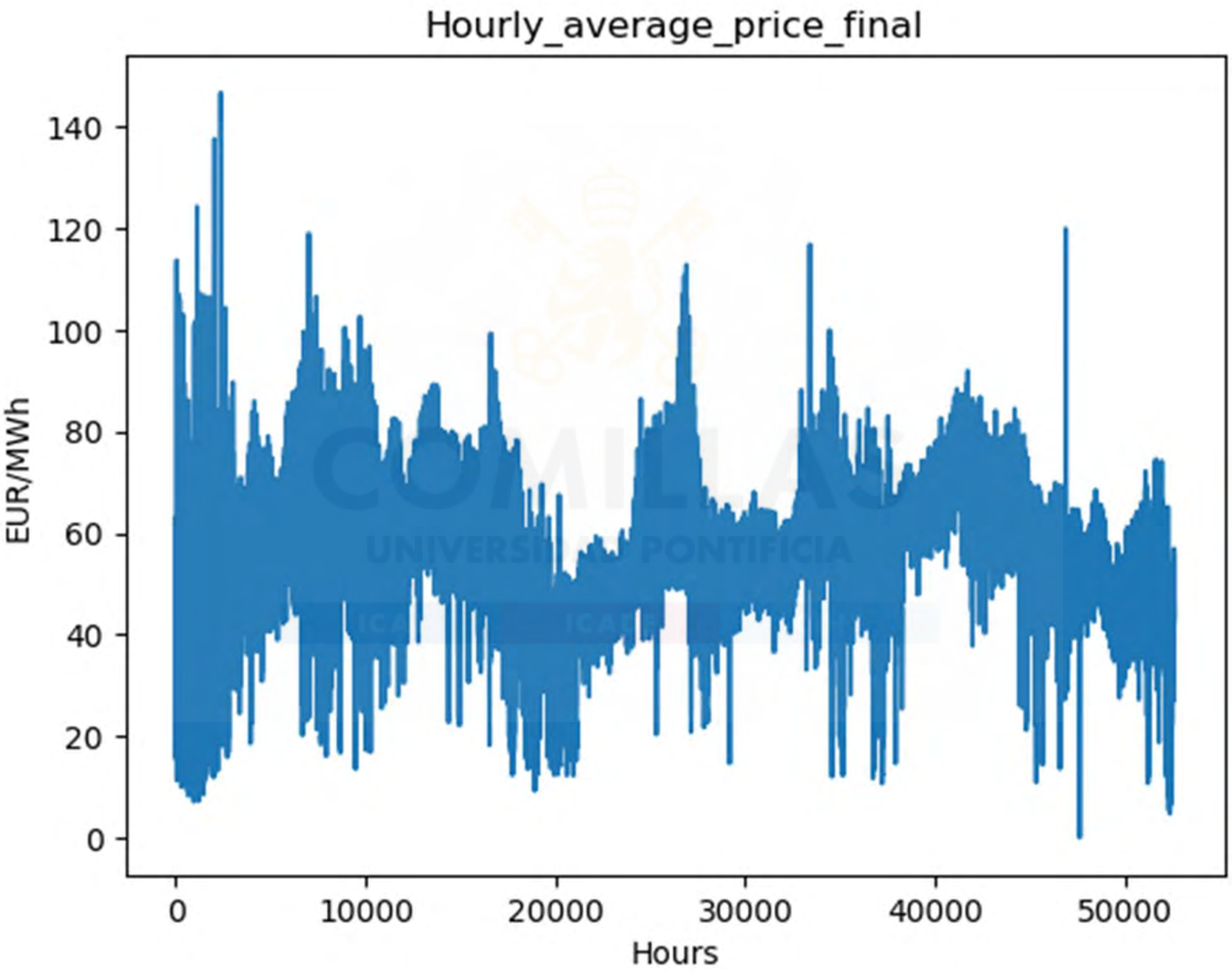
6

Price estimation

Explain the market price of the day-ahead market

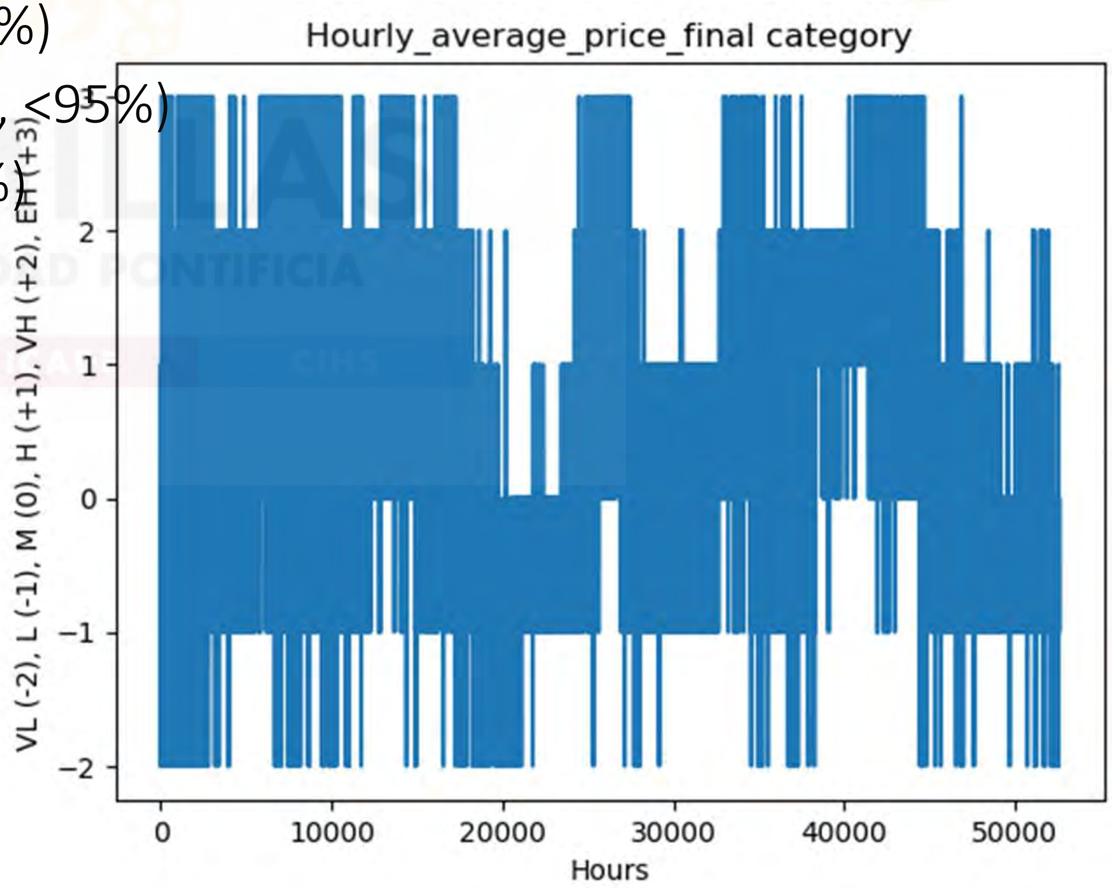
- This price should be related to the output of the technologies

Market price over time



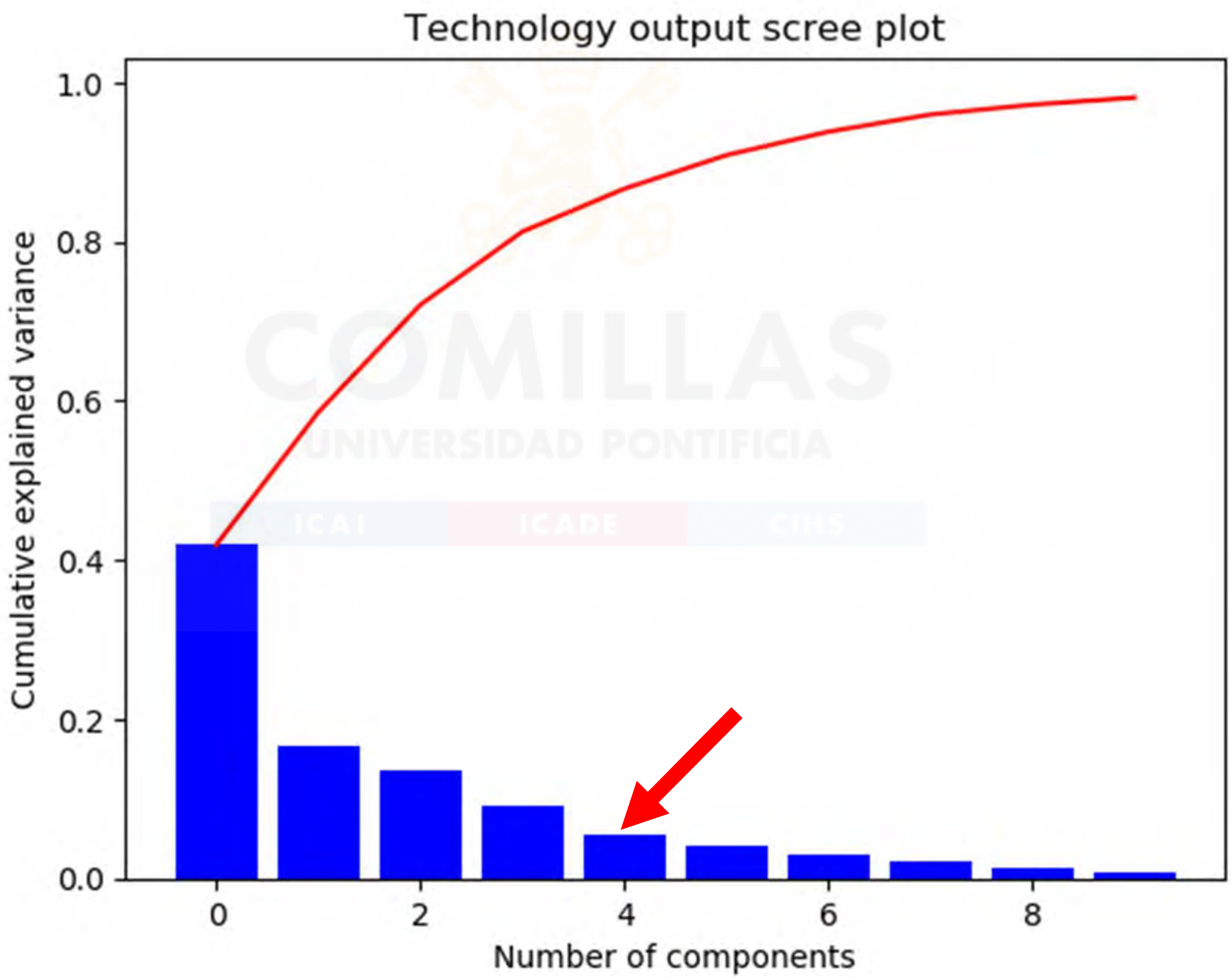
Market price category

- Classification of the prices in categories according to their CDF
  - Very Low (quantile <5%)
  - Low (quantile >5%, <25%)
  - Medium (quantile >25%, <50%)
  - High (quantile >50%, <75%)
  - Very High (quantile >75%, <95%)
  - Extra High (quantile >95%)

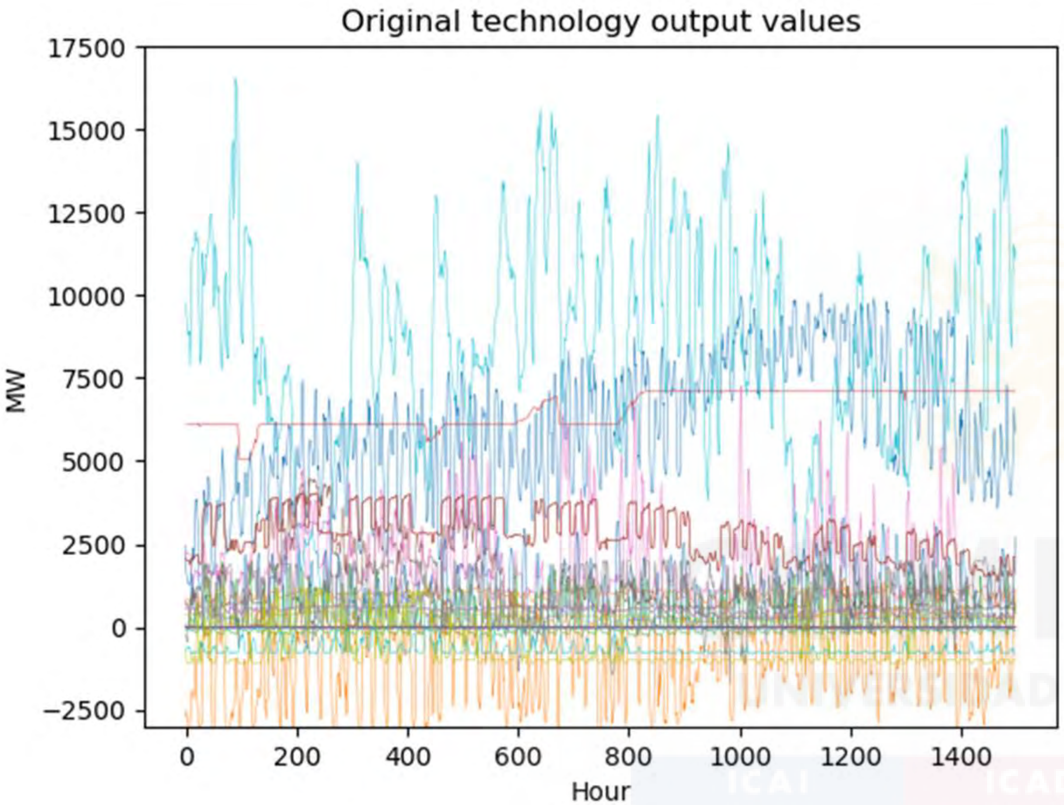


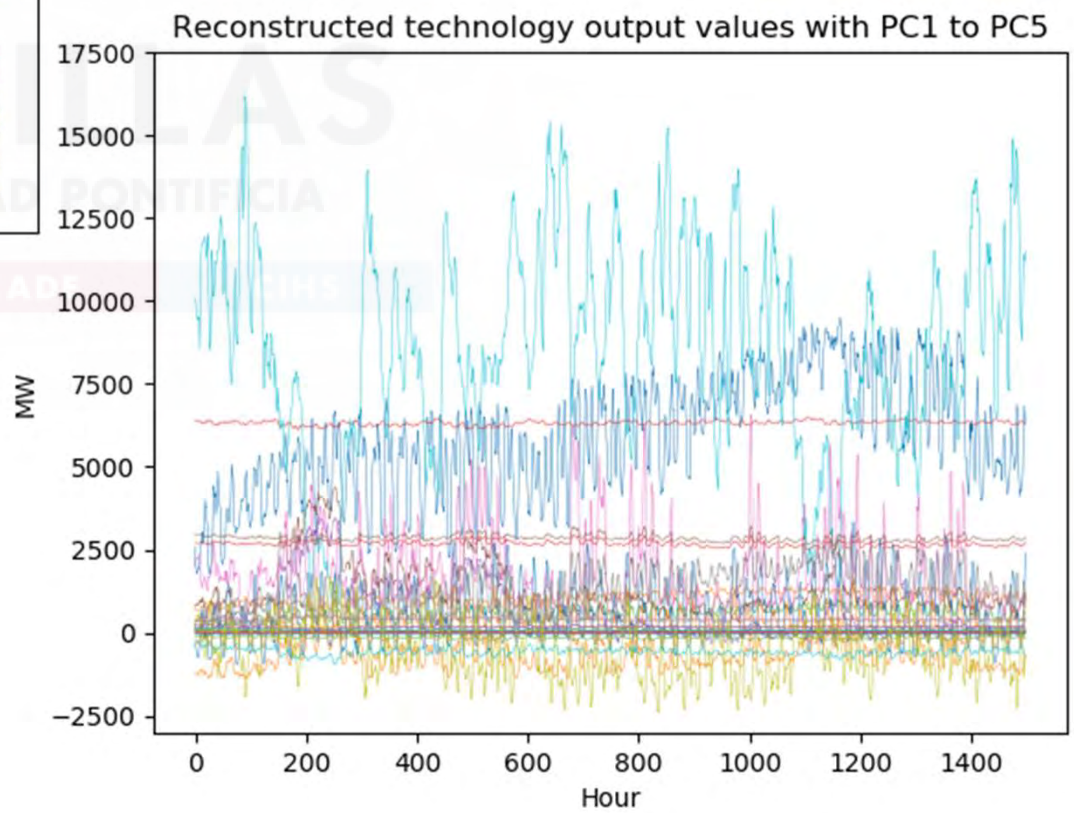
PCA of technology outputs

- PCA is applied to all 31 technology outputs (nuclear, CCGT, solar PV, wind onshore, biomass, coal, etc.)



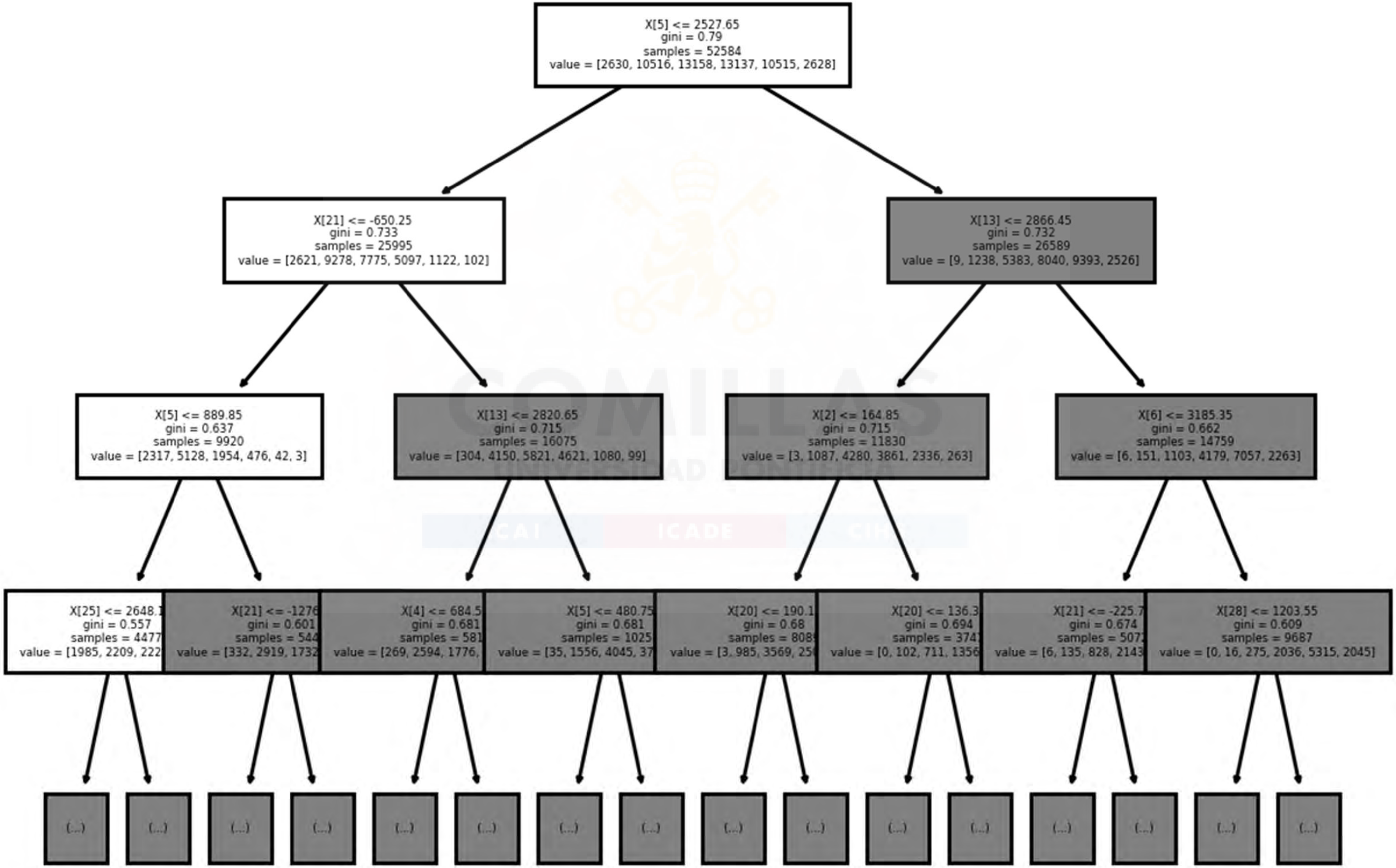
Technology output
Original and reconstructed values with 5 PCs





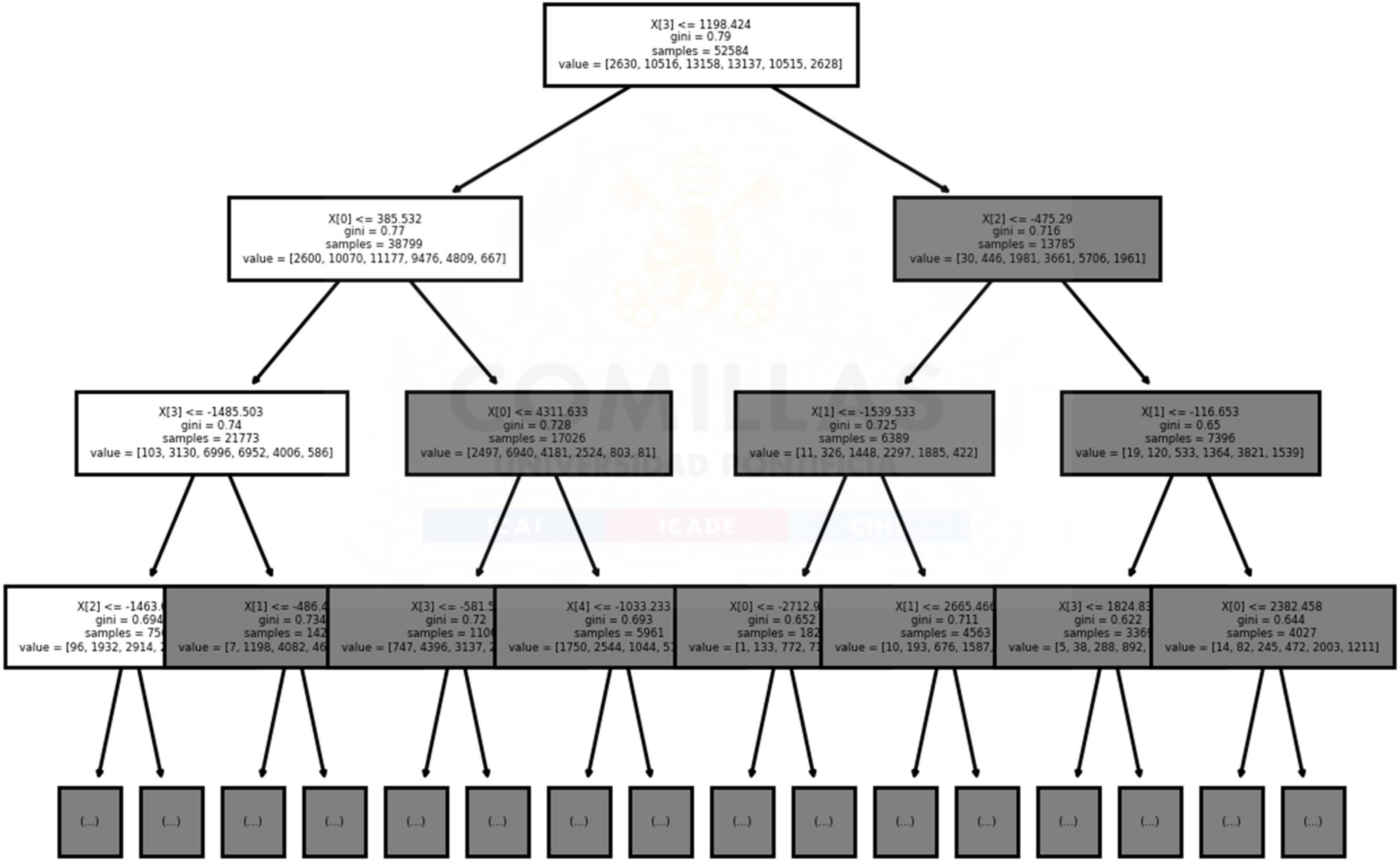
Market price class

Classification tree as a function of tech outputs (31 vars)



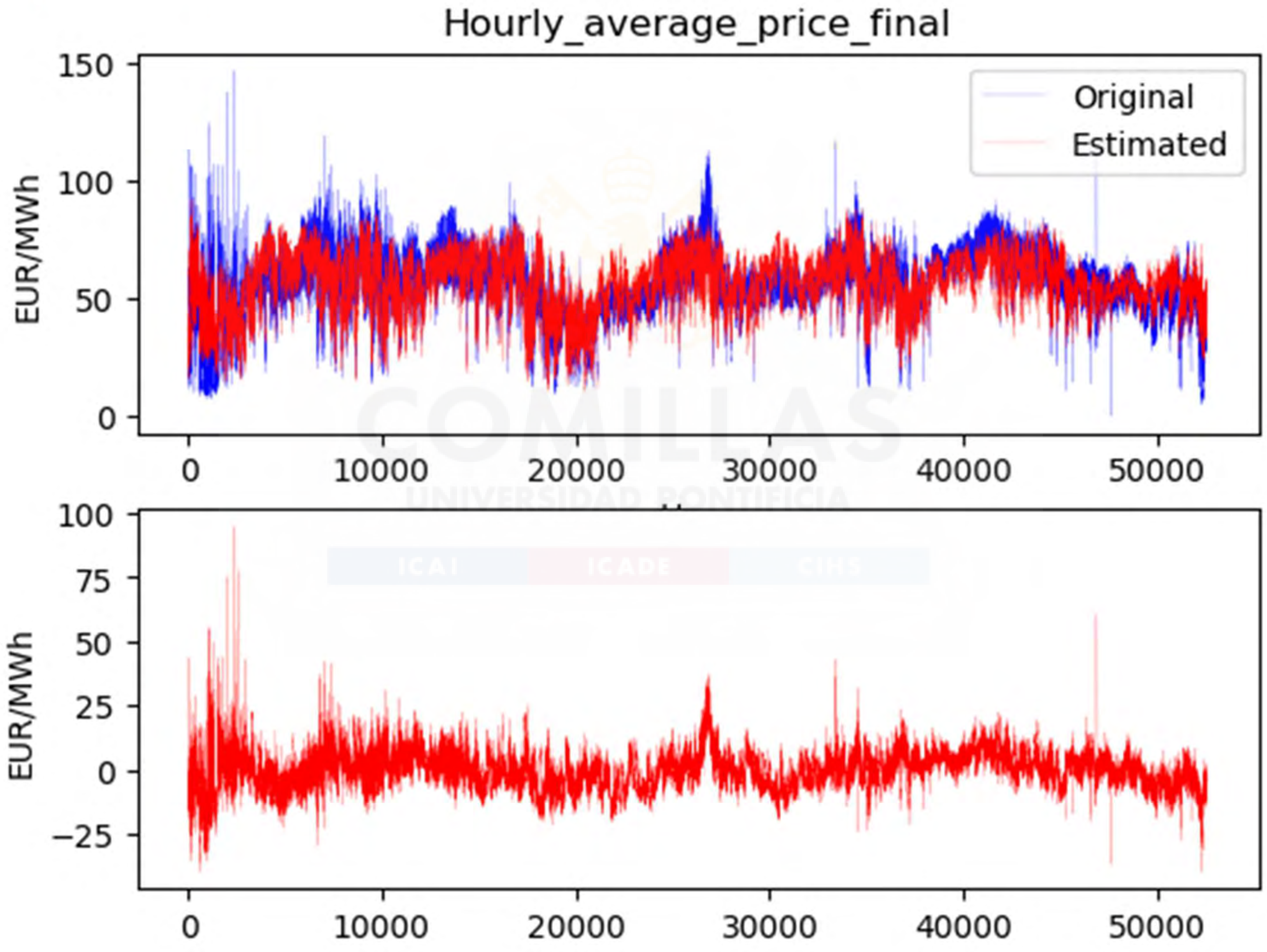
Market price class

Classification tree as a function of PC outputs (5 vars)



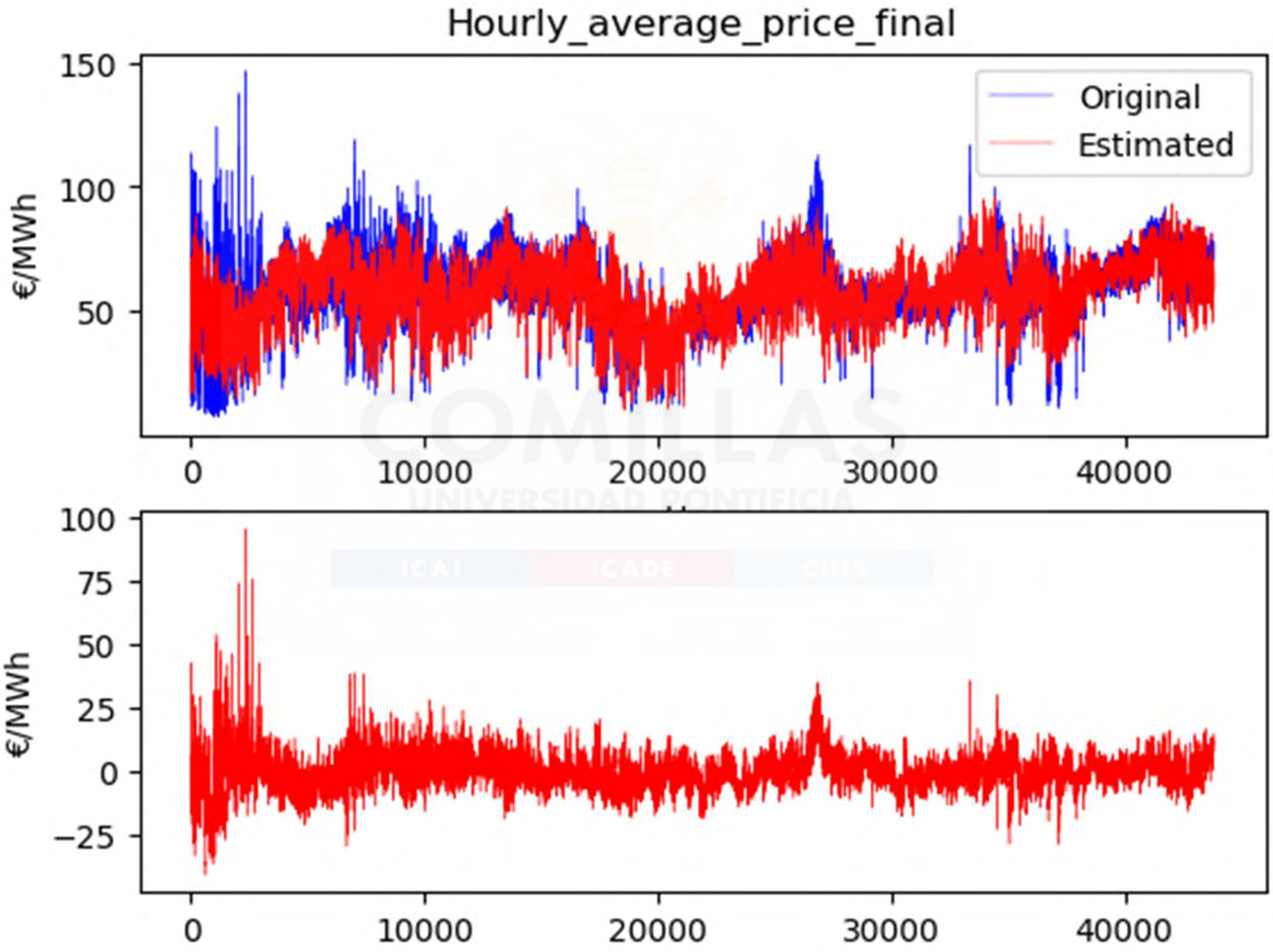
Market price

Original and estimated with linear regression of tech outputs (31 vars)



Market price

Original and estimated with linear regression of PC outputs (5 vars)



Thank you for your attention

Prof. Andres Ramos
https://www.iit.comillas.edu/aramos/
Andres.Ramos@comillas.edu