POLITECNICO DI TORINO

Master’s Degree
in Territorial, Urban, Environmental and Landscape Planning

STATISTICAL MODELLING FOR SIMULATING ELECTRIC ENERGY CONSUMPTION OF RESIDENTIAL USERS IN MENDOZA (AR)

A Master’s Thesis
Submitted to the College of Planning and Design

Author:

Sheref Elsharqawy

Supervisor:

Prof. Guglielmina Mutani

External Supervisor:

Prof. Mariela E. Arboit

Academic Year 2022-2023

I hereby declare that the contents and organization of this dissertation constitute my own original work and do not compromise in any way the rights of third parties, including those relating to the security of personal data.

…………………………………......

Sheref Elsharqawy

Turin, July 11, 2023

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to start with expressing my appreciation and gratitude to INCIHUSA-CONICET represented in Prof. Arboit, for granting me access to the dataset that made complete this research. My deepest gratitude to my supervisor, Prof. Mutani, for her support and guidance throughout my work on the project. Finally, I would like to give a special thanks to my family for their constant source of inspiration, for believing in me and encouraging me to reach my goals, particularly my wife, Ola.

ABSTRACT

The aim of this project is to develop a model that estimates the electrical energy consumption of residential customers in Mendoza's urban region. The research will look at two of the eighteen departments that make up the city of Mendoza. The analytical procedure comprises statistical functions for parameters that have a direct impact on electrical energy consumption, particularly socio-demographic and meteorological factors.

The data was collected from the CONICET-Technological Scientific Centre of Mendoza, and it included consumer electrical bills from 2015 to 2021. The studies performed include data filtering and merging, data validation and clustering, and model creation based on direct-influence factors, using R programming to help cluster the data based on consumption and socio-demographic factors. The reference year of 2016 is used in the procedure, and a model for one department is built and tested on the other. To estimate other users based on the variables listed in each cluster, a multiple regression model equation is employed.

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

1 INTRODUCTION

1.1. ABOUT ELECTRIC ENERGY MODELLING

Due to the rapid growth of consumers in urban areas, electrical consumption has recently been a major source of concern. Electrical consumption modelling aids in calculating user electricity demand for effective energy management and planning. Statistical modelling is a data-driven approach in which input and output factors determine model performance. The data utilized to analyze and build the model was obtained from Mendoza's CONICET center for technological science.

The accuracy of the model is determined by the available data for electrical consumers; the better the built-up data for a bottom-up level, the more accurate the estimation for direct consumption of electricity based on household appliances. However, the data available for the Mendoza database is primarily derived from census sections that span several homes of users, a procedure that necessitates user aggregation and clustering in order to obtain typical consumption types and the accompanying variables from census scale. The goal of the investigation is to identify the closest factors that influence electricity consumption.

1.2. PROJECT SETTING – METHODOLOGY AND STRUCTURE

The key dependent variable in this study is energy consumption (Ec), which is calculated using data from monthly and bimonthly bills received for residential users of the two departments. By considering census-scale socio-demographic parameters as well as climatic conditions to determine the strongest link with electricity usage, and by following a flowchart of data collection and definition, pre-processing of filtering and grouping.

The clustering analysis is carried out twice, the first time with the typical customers in terms of absolute family yearly spending, and the second time with the inclusion of social-demographic characteristics in the clustering. The pre-processing output and selected variables are then tested using linear and multiple regression analysis. The final stage in validating the model coefficients on the geographical level is to divide the data into two datasets, discover the equation on one department, and then test the result on the other.

1.3. RESEARCH QUESTIONS

a. What factors influence the electrical energy consumption for the residential users?

b. What is the composition of typical users in terms of electricity consumption in the study area of Mendoza?

c. What are the benefits from electrical energy modelling to the future of the city?

1.4. LIMITATION OF THE STUDY

According to the literature, the most proven and accurate method for estimating and predicting energy consumption, particularly for electrical energy consumption, is the bottom-up approach that collects direct end-user components composed of user behavior and household appliances. The elements that directly impact electricity consumption are not covered in this study, but such information can be obtained by smart metering or a random survey of a sample of consumers.

Data on how and when building energy demand changes throughout the day for various end uses such as appliances, lights, ventilation, heating, and cooling are used to create energy load profiles (Luo et al., 2017; Zakovorotnyi & Seerig, 2017). The available data is used as averages retrieved from the analysis of census section factors, which are then redistributed across users assuming that users exhibit comparable behavior for each census section.

2 LITERATURE REVIEW

2.1. INTRODUCTION

Several methods and approaches for energy modelling for the residential use sector, which accounts for a substantial fraction of global energy consumption, have been studied in the literature. In general, the fundamental role of these approaches is to provide accurate estimates of energy consumption, which can assist policymakers and energy suppliers in making the best decisions for green building design, energy planning, and emission reduction initiatives. The analysis of energy consumption normally accounts for all residential energy types, such as electricity, domestic hot water (DHW), space heating (gas or electric), and space cooling.

2.2. ENERGY MODELLING

Swan and Ugursal (2009) discuss two main approaches to energy modelling in their paper *"Modelling of End-Use Energy Consumption in the Residential Sector: A Review of Modelling Techniques,"* which branches out generally to top-down and bottom-up modelling, showing the methods used for each approach and a review of previous analyses used in the models associated (**Figure 1**). They classified the two approaches as follows: Top-down consists of econometric and technological methods, where the main idea is to deal with energy as a sink without focusing on actual end-user consumption and relying on econometric variables of interest that affect energy consumption. Bottom-up approaches, on the other hand, are divided into two major categories: *Statistical* and *Engineering*, and each has a variety of energy modelling methodologies.

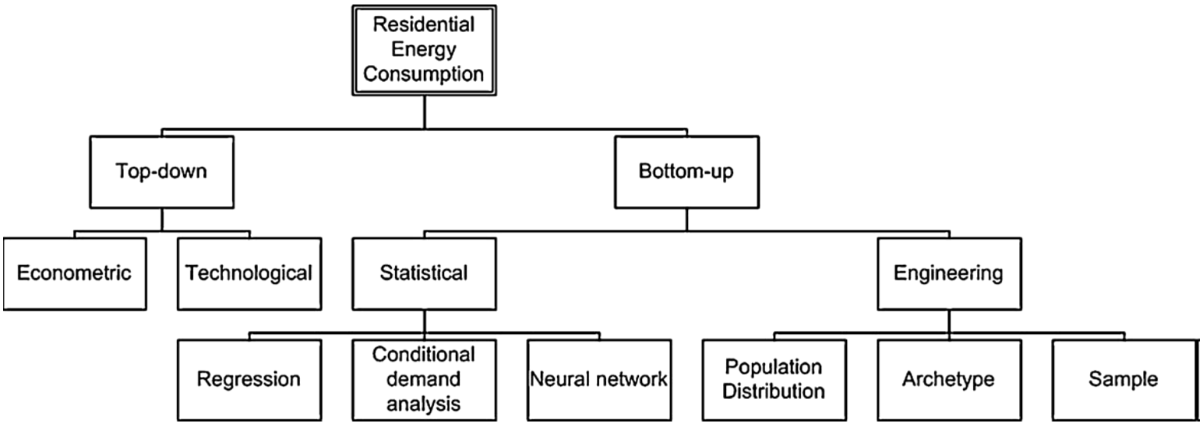
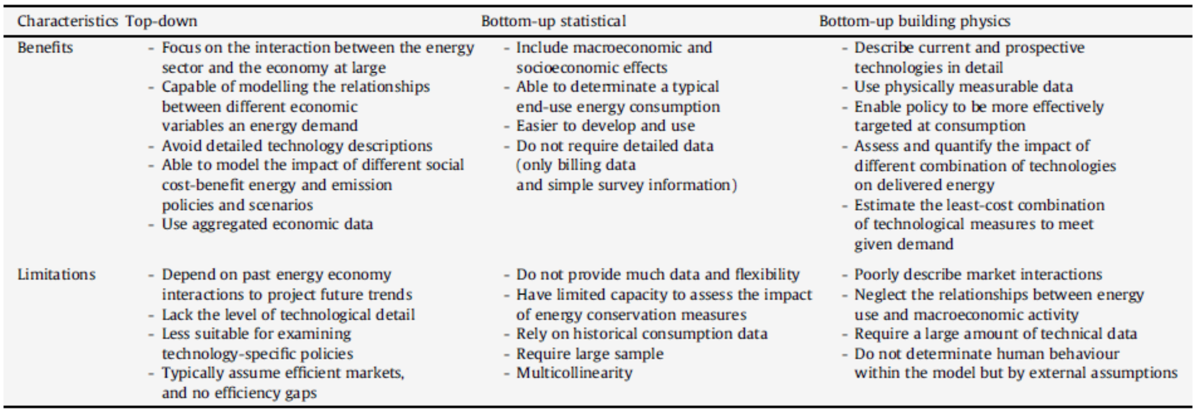


Figure 1. Top-Down and Bottom-Up Modeling Techniques for Estimating the Regional or National Residential Energy Consumption. ***Source:*** Swan & Ugursal (2009)

The fundamental premise of bottom-up modelling is to regress the link between end-uses and overall consumption based on a population sample size that is representative of the entire sample. Statistical modelling (SM) is based on end-use, house, and user behaviour characteristics, with the sample of houses used undergoing one of the three well-documented techniques mentioned: *Regression*, *Conditional Demand Analysis (CDA),* and *Neural Networks (NN),* all of which contribute to the same goal.

Engineering modelling (EM) is based on the physical properties of end-use appliances and may or may not refer to previous consumptions. It employs three primary techniques: *Distributions*, *Archetypes*, and *Samples*. The main idea for the three techniques is to calculate the energy consumption of the end-use for a sample of the population, then scale up to represent a region.

Following that, in 2010, Kavgic et al. conducted a study that discussed the benefits and limits of the two main methodologies used in energy modelling, top-down and bottom-up (**Table 1**), as well as the data flow between the two views, depending on regression approaches, statistical modelling, such as the PRISM scorekeeping method at Princeton, US, with a year of monthly energy billing and linear regression analysis.



Characteristics	Top-down	Bottom-up statistical	Bottom-up building physics
Benefits	- Focus on the interaction between the energy sector and the economy at large - Capable of modelling the relationships between different economic variables an energy demand - Avoid detailed technology descriptions - Able to model the impact of different social cost-benefit energy and emission policies and scenarios - Use aggregated economic data	- Include macroeconomic and socioeconomic effects - Able to determinate a typical end-use energy consumption - Easier to develop and use - Do not require detailed data (only billing data and simple survey information)	- Describe current and prospective technologies in detail - Use physically measurable data - Enable policy to be more effectively targeted at consumption - Assess and quantify the impact of different combination of technologies on delivered energy - Estimate the least-cost combination of technological measures to meet given demand
Limitations	- Depend on past energy economy interactions to project future trends - Lack the level of technological detail - Less suitable for examining technology-specific policies - Typically assume efficient markets, and no efficiency gaps	- Do not provide much data and flexibility - Have limited capacity to assess the impact of energy conservation measures - Rely on historical consumption data - Require large sample - Multicollinearity	- Poorly describe market interactions - Neglect the relationships between energy use and macroeconomic activity - Require a large amount of technical data - Do not determinate human behaviour within the model but by external assumptions

Table 1. Benefits and Limitations of Bottom-Up and Top-Down Modelling Approaches. ***Source:*** Kavgic et al. (2010)

The authors then carried on with a comparison of nine bottom-up models revealing varied dimensions of data intake and output, as well as the scale of data size and aggregation level. The structure and data flowchart of five physics-based energy models used in the UK are then discussed to show the advantages and disadvantages of various methods for the same building stock of housing. The purpose of this comparison is to clarify the difference between, and the limitation of using, these physical models, showing the contribution of such models in describing energy losses and technological effects, and include relationships between the effects in a way that could be used for energy consumption estimation models.

Fumo's 2014 paper discusses the fundamentals of general building energy modelling without focusing on any specific approach. The study provides an evaluation and credibility score for various modelling approaches based on suitability criteria for new building design, energy saving measures, and retrofit analyses. The author then describes calibration in terms of fitting the model to the data obtained and emphasises the significance of the verification process, which accepts simulation findings when compared to the reference data. Weather data is demonstrated to be important for the building energy modelling process, where meteorological characteristics play a significant role in the variation of energy consumption on the level of space heating and cooling. As most of the time, the actual data cannot be directly considered in the analysis or inserted into a model or software without a pre-step of correcting or normalising the data, the process of inserting weather data and regarding the parameters should be done with great care.

In a 2015 study conducted in Singapore by Chuan and Ukil, the authors demonstrated the use of a bottom-up approach in the process of energy modelling as well as energy load profiling and then applied a mathematical model that represents the case study of Singapore and presents the residential stock and how it is classified into five types based on the number of rooms per dwelling.

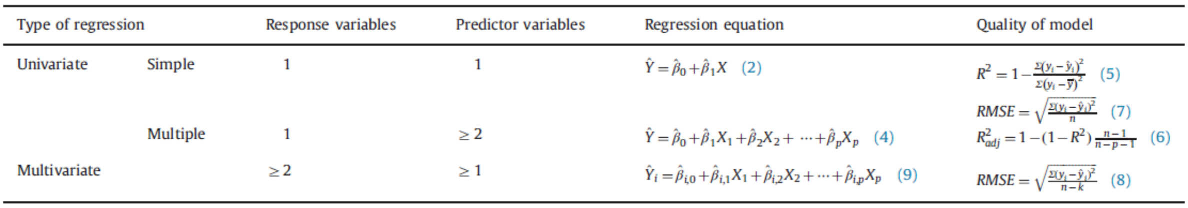
The data is next checked against the planned load profiling, utilising the sum of the appliance usage curve to reflect the household load curve based on the actual observed power consumption. The mathematical model utilised is based on the simulation of a load curve generation loop for appliances, taking into account the starting probability, appliance saturation, and a statistical probability factor. After that, the simulation process is applied to the five categories of households based on a probability factor, and the simulated consumption is compared and validated against the real consumption data collected in a case study of Nanyang Technological University campus housing.

During the same year, 2015, Subba et al. published a conference paper that analysed monthly billing data from roughly 70 families in the kingdom of Bhutan to identify the most linked variables that affect electrical energy use. The study focused on three major socio-demographic factors: dwelling occupancy, home size, and occupant income level. The monthly annual consumption trend showed a significant association with the factors listed; however, the seasonal effect of temperature could not be confirmed.

Then, in 2020, Lamagna et al. introduced the process of obtaining electrical energy hourly load profiles from Procida city hall monthly billing data, taking into consideration the structure of the Italian billing system, which is separated into three periods of F1-F2-F3 that indicate peak, mid-level, and off-peak use across a weekly time frame. Following that, a normalisation is performed between the known profile and the data obtained to arrive at a simulation for weekday consumption of the city hall case, and the findings are validated by employing error analysis between the anticipated and real hourly load profiles.

Another notable contribution by Alqasim (2022) was the use of multiple linear regression using ARIMA to predict energy usage, which he applied to a case study of 27 Dubai police facilities. The thesis describes the entire modelling process and data flowchart, beginning with the initial phase of data collection and continuing through data processing to clearly prepare the dataset free of anomalies and ensure the data quality dimensions become valid for subsequent processes. A correlation matrix is used to further describe the link between included variables after expanding the information with numerous factors of relevance, such as the rank area categorization for police facilities, the monthly mean temperature, and seasonal months of bills. Using monthly bills for power and water consumption, the dataset is split into 70% for training and the remaining 30% for validation and testing. A machine learning model capable of predicting energy usage is then built using the Multiple Linear Regression (MLR) model, and another model, Autoregressive Integrated Moving Average (ARIMA), is applied. When the two models were compared, MLR performed better in terms of accuracy than ARIMA.

The most important paper linked to this research is Fumo and Rafe Biswas's 2015 study, *"Regression Analysis for Prediction of Residential Energy Consumption."* The author alludes to the fundamentals of regression modelling approaches used in the field of home energy consumption prediction in this study. The relation ruling regression analysis is intended to classify the type of analysis performed, with the number of response variables separating univariate and multivariate analysis. The distinction between single linear and multiple linear regression in linear regression is the use of single or multiple independent variables in the analysis process of predicting the dependent variable (**Table 2**). To quantify the quality of model fitting, some coefficients such as the R2 determination coefficient and the Root Mean Square Error (RMSE) are used.



Type of regression		Response variables	Predictor variables	Regression equation	Quality of model
Univariate	Simple	1	1	$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X$ (2)	$R^2=1-\frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\overline{y})^2}$ (5) $RMSE=\sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n}}$ (7)
	Multiple	1	≥ 2	$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\cdots+\hat{\beta}_p X_p$ (4)	$R^2_{adj}=1-(1-R^2)\frac{n-1}{n-p-1}$ (6)
Multivariate		≥ 2	≥ 1	$\hat{Y}_i=\hat{\beta}_{i,0}+\hat{\beta}_{i,1} X_1+\hat{\beta}_{i,2} X_2+\cdots+\hat{\beta}_{i,p} X_p$ (9)	$RMSE=\sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n-k}}$ (8)

Table 2. Classification of Linear Regression Approaches. ***Source:*** Fumo & Rafe Biswas (2015)

Collinearity is a phrase that describes model instability. It occurs when there is a correlation between the independent variables that comprise the model, a problem that can be resolved by employing Principle Component Analysis (PCA). The phrase "energy signature" refers to the climatic effect on building energy consumption, and the least squares regression method is preferred. The author's viewpoint emphasises the need for applying energy modelling on an individual scale to improve prediction accuracy rather than using a sample of the population. The research advises using a quadratic regression model to capture the influence of HVAC systems only when data is collected on an hourly basis.

2.3. CONCLUSION

From prior work on the topic of energy modelling, there are fundamentals and components of concern that should be followed when processing the dataset of the project of interest. The modelling phase necessitates pre-processing activities to prepare the data for testing and validation, including filtering and cleaning, normalization, aggregation, and correlation approaches. When it comes to the dimension of data available and the desired output that must be tested, the input data type could form and suggest the workflow of the entire procedure. Regression analysis is a sort of statistical modelling in which the relationship and equation for the inter-related variables are directly proposed, followed by statistical tests for error that emerge as a result of the difference between real and projected data.

3 PROJECT DESCRIPTION

3.1 INTRODUCTION TO CASE STUDY

Mendoza is the capital of the province of the same name in Argentina. The province, which is located in the country's central west (refer to **Figure 2**), is divided into 18 departments and has a population of around 2 million people, according to the general census of 2022 (see **Figure 4** for the population evolution of Mendoza up to 2001). The scope of this study is limited to two central departments, the Capital and Godoy-Cruz (**Figure 3**), which have a combined population of around 300,000 people and cover an area of 160 km^2. Total energy use in the residential sector accounts for approximately 37.8% and 30.8% for space heating and hot water, respectively, followed by 0.9% and 11.7% for lighting, ventilation, and other electrical appliances. Notably, overall electric consumption in terms of energy sources accounts for around 22.6% (Reta, 2012).

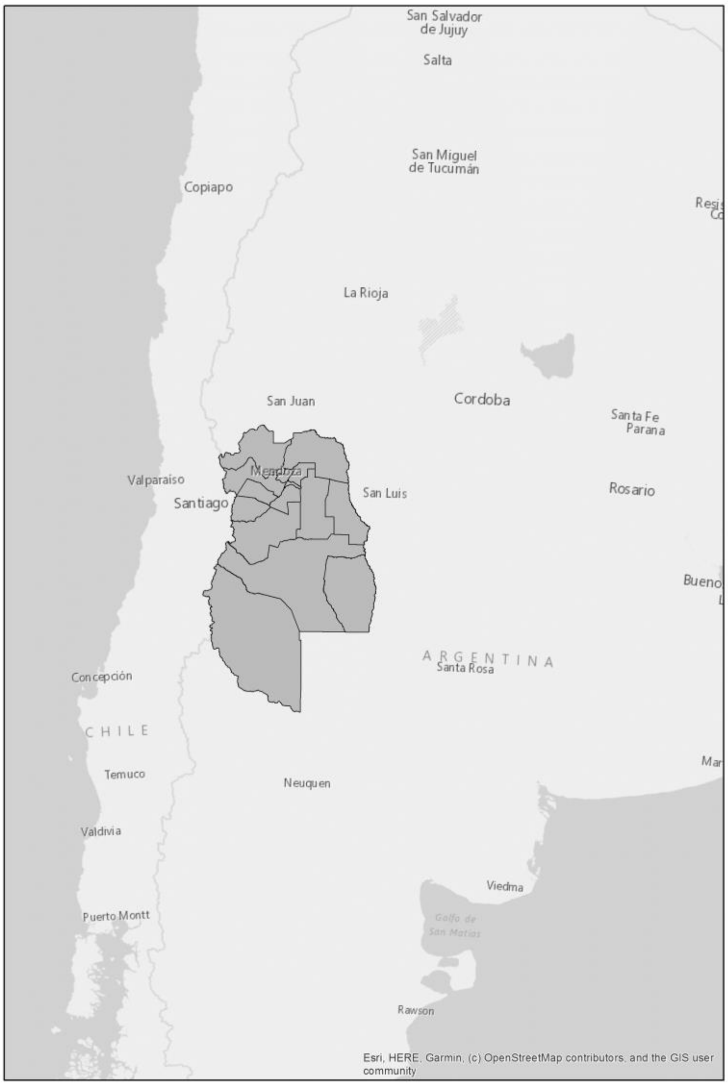


Figure 2. Mendoza Province's Location in Argentine. ***Source:*** Author

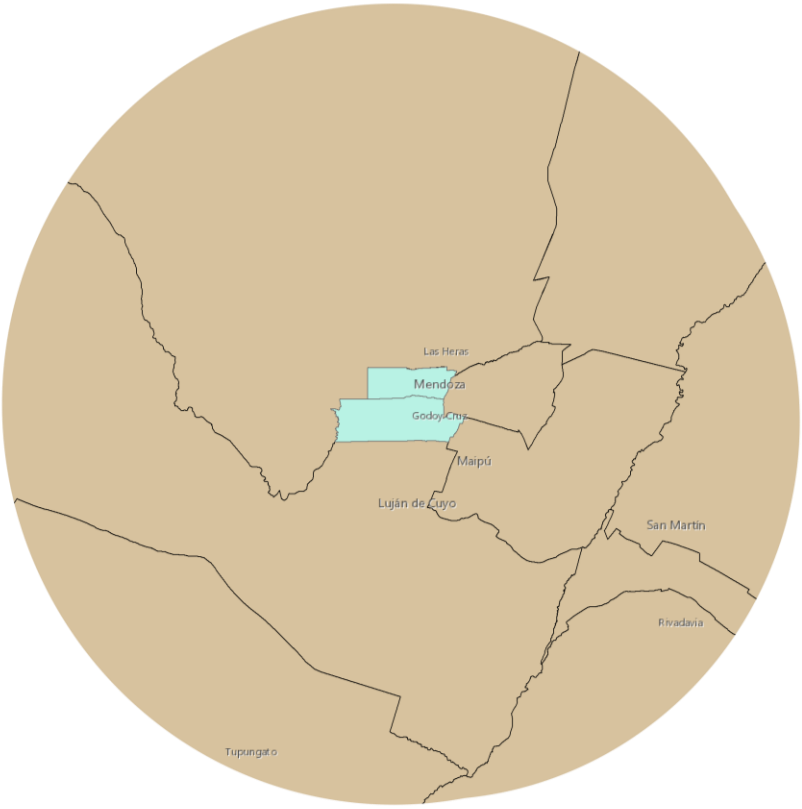


Figure 3. Study Area of the Two Departments: Capital and Godoy Cruz. ***Source:*** Author

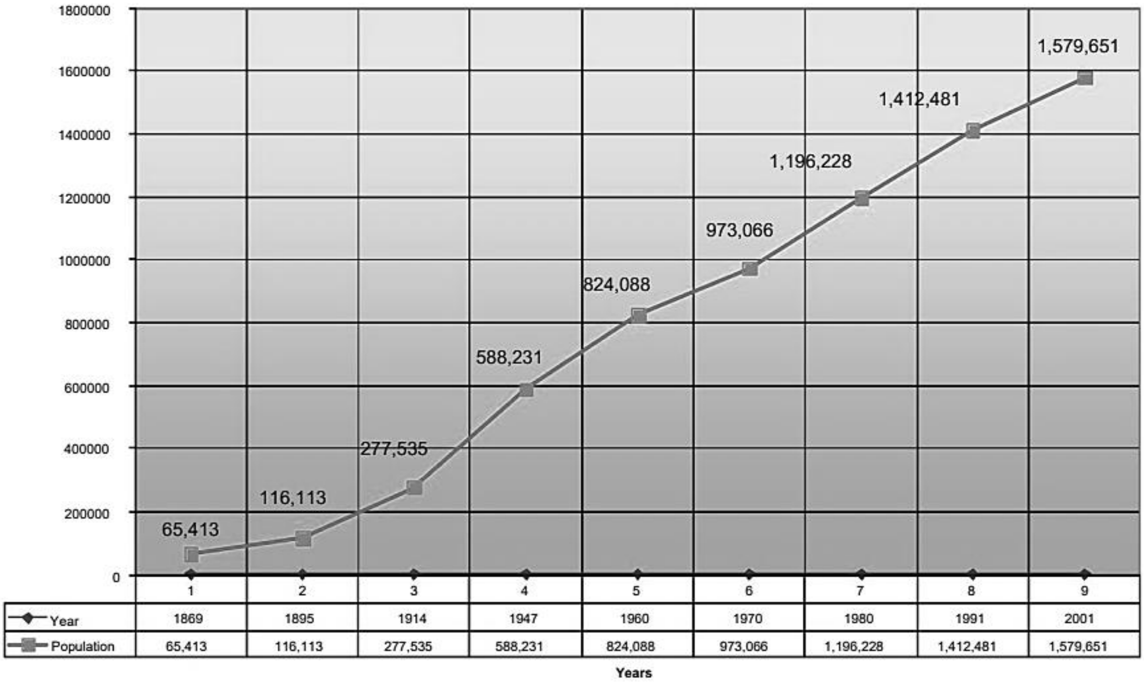


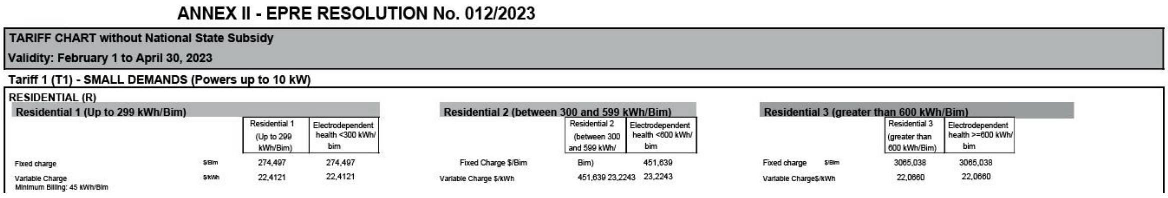
Figure 4. Evolution of Population of Mendoza Province 1869-2001. ***Source:*** DEIE-Mendoza

3.2. DATA SOURCE

The dataset developed for the study region is made up of six main databases, which are merged and aggregated to exactly incorporate all available data of concern. The main source of data is CONICET-Mendoza, which is a technological and scientific facility that provides users with data on electrical monthly and bimonthly bills for seven reference years, as well as cadastral, census, gas model, and GIS databases, while the author adds the weather data.

3.3. DATA COLLECTION

The data for user energy bills is collected by the provincial electrical regulated authority (EPRE) by collecting metered use from electrical energy supply firms that work in the relevant departments. The historical bill database spans seven reference years, from 2015 to 2021, with around 106,000 registered users, with monthly and bimonthly consumption bills recorded for each user with associated (X, Y) coordinates. Providers adhere to the general Tariff category (**Figure 5**) established by the government and place their terms of reference locally, which show the categories of consumers based on consumption range, the range that is used for subsidy assignment.



ANNEX II - EPRE RESOLUTION No. 012/2023

TARIFF CHART without National State Subsidy
Validity: February 1 to April 30, 2023

Tariff 1 (T1) - SMALL DEMANDS (Powers up to 10 kW)

RESIDENTIAL (R)

Residential 1 (Up to 299 kWh/Bim)

		Residential 1 (Up to 299 kWh/Bim)	Electrodependent health <300 kWh/ bim
Fixed charge	$/Bim	274,497	274,497
Variable Charge Minimum Billing: 45 kWh/Bim	$/kWh	22,4121	22,4121

Residential 2 (between 300 and 599 kWh/Bim)

	Residential 2 (between 300 and 599 kWh/	Electrodependent health <600 kWh/ bim
Fixed Charge $/Bim	Bim)	451,639
Variable Charge $/kWh	451,639 23,2243	23,2243

Residential 3 (greater than 600 kWh/Bim)

		Residential 3 (greater than 600 kWh/Bim)	Electrodependent health >=600 kWh/ bim
Fixed charge	$/Bim	3065,038	3065,038
Variable Charge$/kWh		22,0660	22,0660

Figure 5. Tariff Composition Provided by Distributor. ***Source:*** G_Coop. Eléctrica Godoy Cruz

4 DATA ANALYSIS AND MODELLING

4.1. DATA PREPARATION AND WORKFLOW

Following the receipt of databases, the pre-modelling process begins by studying the data provided in order to gain a better understanding of the data type and search for the documented definition. Then, to avoid data inconsistency, a data cleaning and filtering process is carried out to consider and delete any incorrectly added data or abnormalities. As indicated in the image below, the data pipeline is multi-directional, with distinct databases depicted independently one by one during the data definition phase (**Figure 6**).

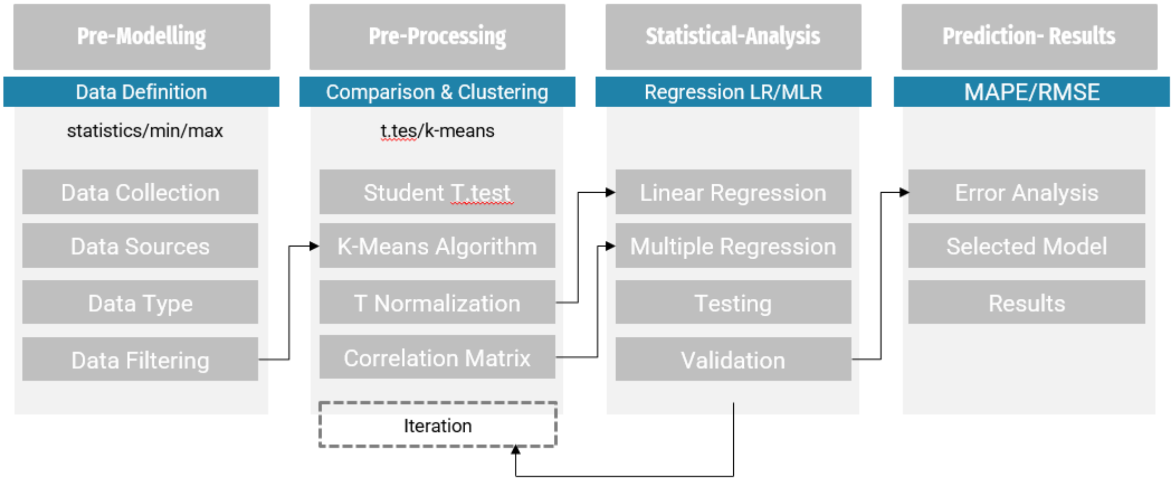
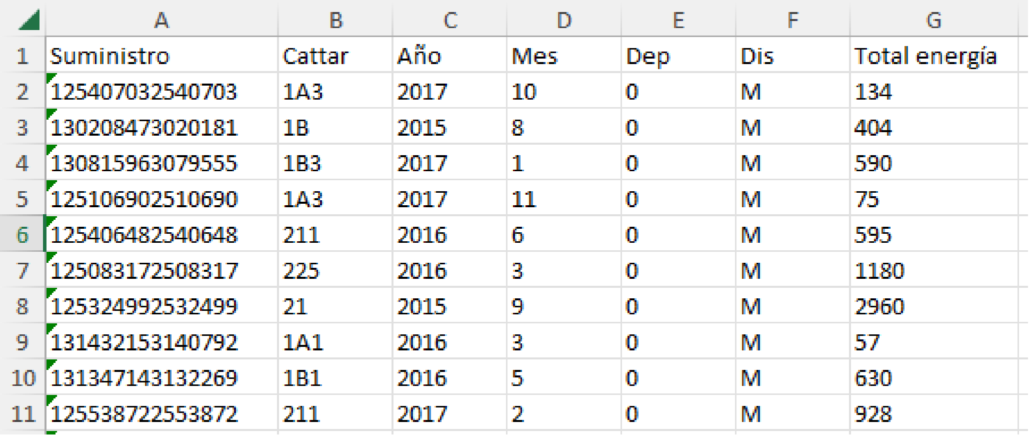


Figure 6. Data Workflow of the Study Model. ***Source:*** Author

4.1.1. Electricity Consumption (Ec)

When aggregated in terms of user-ID (Suministro) to consider all the recorded bills for the distinct user, the user's consumption raw data shows the recorded bills of 12-month and 6-bimonth for the annual consumption (**Figure 7**), with the total number of users of the two departments accounting for about 114,840 users. The reference years of records are from 2015 to 2021, with the majority of aggregated bills count per user are in bimonthly time series, as compatible with the general billing system of Mendoza province.



	A	B	C	D	E	F	G
1	Suministro	Cattar	Año	Mes	Dep	Dis	Total energía
2	125407032540703	1A3	2017	10	0	M	134
3	130208473020181	1B	2015	8	0	M	404
4	130815963079555	1B3	2017	1	0	M	590
5	125106902510690	1A3	2017	11	0	M	75
6	125406482540648	211	2016	6	0	M	595
7	125083172508317	225	2016	3	0	M	1180
8	125324992532499	21	2015	9	0	M	2960
9	131432153140792	1A1	2016	3	0	M	57
10	131347143132269	1B1	2016	5	0	M	630
11	125538722553872	211	2017	2	0	M	928

Figure 7. Sample of Recorded Bills in the Raw Database of Ec. ***Source:*** Author

All Ec databases are aggregated in reference years to prepare for the pre-processing step, and anomalies are detected for negative bills that are assigned as a deposit for future payment but do not reflect the real consumption of the user, and zero bills that refer to temporary use of a dwelling when referring to permanent users of annual consumption (refer to **Table 3**).

Year	2015	2016	2017	2018	2019	2020	2021
Total_User_Count	110,923	113,148	113,176	114,773	116,638	116,965	118,276
Total_Ec_(kWh/Y)	333,603,096	327,479,786	319,392,155	294,848,773	280,530,766	299,038,470	282,911,462
Min Record (kWh/y)	-5,123	-2,840	-14,005	-10,558	-40,573	-16,678	-10,942
Max Record (kWh/y)	999,990	91,816	154,902	68,898	133,155	141,512	146,164
Avg Record (kWh/y)	3,008	2,894	2,822	2,569	2,405	2,557	2,392
Median Record (kWh/y)	2,475	2,365	2,293	2,073	1,933	2,073	1,902
Avg-Med Diff	533	529	529	496	472	484	490
St.dev(P) Record (kWh/y)	3,890	2,468	2,503	2,301	2,239	2,333	2,282
Variance (st.dev^2)	15,135,463	6,093,385	6,264,260	5,295,457	5,011,034	5,442,239	5,207,910
Ec_majority_(avg+Stdev)	6,898	5,363	5,325	4,870	4,644	4,890	4,674
No_Users	105,027	101,666	102,299	103,652	105,576	105,448	107,228
%_Users	94.7%	89.9%	90.4%	90.3%	90.5%	90.2%	90.7%

Table 3. Aggregated Bills for Capital and Godoy Cruz Departments with Statistical Definitions. ***Source:*** Author

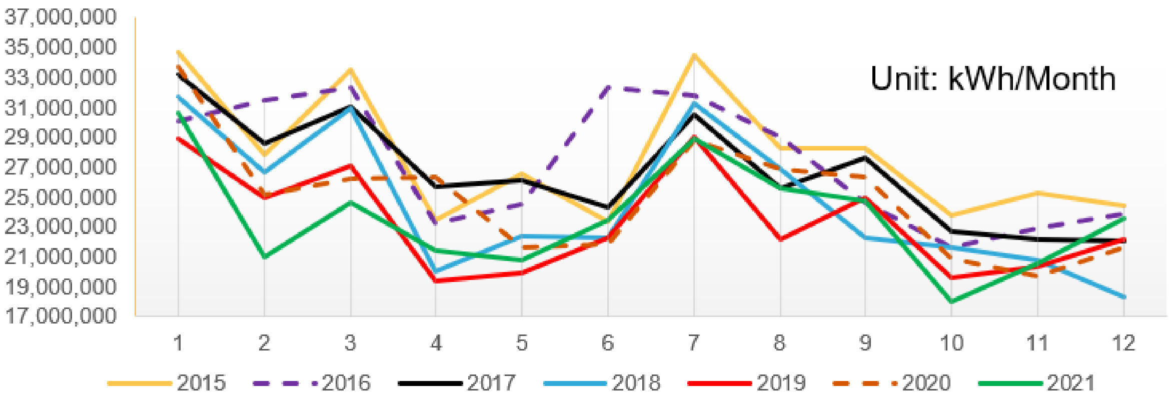


Figure 8. Monthly Annual Total Consumption by Reference Year 2015-2021. ***Source:*** Author

The relevance of the consumption trend occurs in the increase in user count versus the drop in total consumption from 2015 to 2021, while the average yearly consumption shows a decreasing trend, indicating a shift in user behavior or the technological influence of energy-saving appliances (see **Figure 8**).

Anomalies in the Ec database include users with bill counts surpassing 12 months, the number of some users reaching 30 bills, which is not logically inserted, and the removal of users with an additional 12 aggregated bills (**Table 4**).

Year	2015	2016	2017	2018	2019	2020	2021
Total no_Bills(records)	891,814	921,811	926,843	935,783	962,733	970,408	966,996
Min Bills Count	1	1	1	1	1	1	1
Max Bills Count	12	12	12	12	12	12	12
Avg Bills Count	8	8	8	8	8	8	8
Median Bills Count	6	7	6	6	6	6	6
St.dev Count	3	3	3	3	3	3	3

Table 4. Users Record of Bills Over Reference Year 2015-2021. ***Source:*** Author

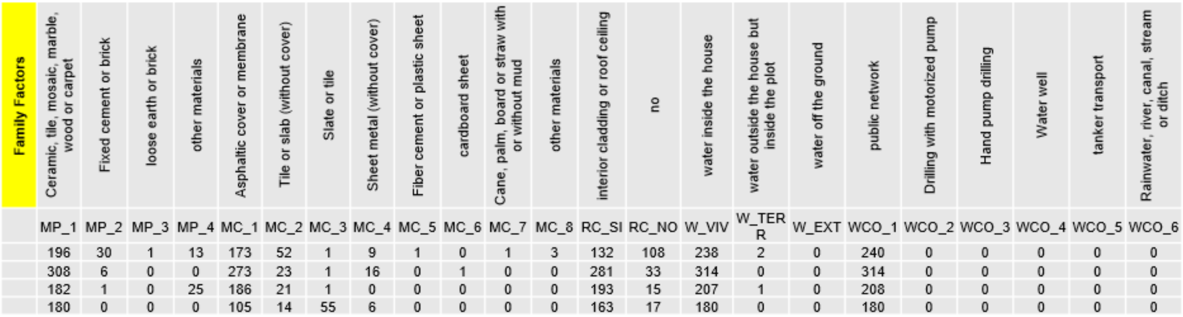
4.1.2. Georeferencing

The data is aligned and georeferenced using ArcGIS software to allocate buildings, census sections, and user layers for analysis. The base coordinates of the data are locally assigned to the Argentine POSGAR 94, zone number 2, and the data is then projected to the world UTM WGS 84, Z19S.

4.1.3. Census Sections

The data for socio-demographic indicators from the reference year 2010 are divided into three major categories: families, households, and population factors. As illustrated in the image below, census variables are recorded on the census section scale for several buildings with similar features.

The characteristics of which properties are indicated in codes, as well as the relative translation of each class category (**Figure 9**). The codes are linked to the appropriate census component.



Family Factors	Ceramic, tile, mosaic, marble, wood or carpet	Fixed cement or brick	loose earth or brick	other materials	Asphaltic cover or membrane	Tile or slab (without cover)	Slate or tile	Sheet metal (without cover)	Fiber cement or plastic sheet	cardboard sheet	Cane, palm, board or straw with or without mud	other materials	interior cladding or roof ceiling	no	water inside the house	water outside the house but inside the plot	water off the ground	public network	Drilling with motorized pump	Hand pump drilling	Water well	tanker transport	Rainwater, river, canal, stream or ditch
	MP_1	MP_2	MP_3	MP_4	MC_1	MC_2	MC_3	MC_4	MC_5	MC_6	MC_7	MC_8	RC_SI	RC_NO	W_VIV	W_TERR	W_EXT	WCO_1	WCO_2	WCO_3	WCO_4	WCO_5	WCO_6
	196	30	1	13	173	52	1	9	1	0	1	3	132	108	238	2	0	240	0	0	0	0	0
	308	6	0	0	273	23	1	16	0	1	0	0	281	33	314	0	0	314	0	0	0	0	0
	182	1	0	25	186	21	1	0	0	0	0	0	193	15	207	1	0	208	0	0	0	0	0
	180	0	0	0	105	14	55	6	0	0	0	0	163	17	180	0	0	180	0	0	0	0	0

Figure 9. Sample Of Census Characteristics. ***Source:*** Author

4.1.4. Buildings Data

The spatial allocation of the building shapefile indicates exclusively residential use in the two departments, with the attributes indicated being the shape area and floor number of each feature, and the graph below illustrates the count of buildings based on floor numbers (**Figure 11**). And the following map shows the distribution of buildings according to the number of floors (**Figure 10**).

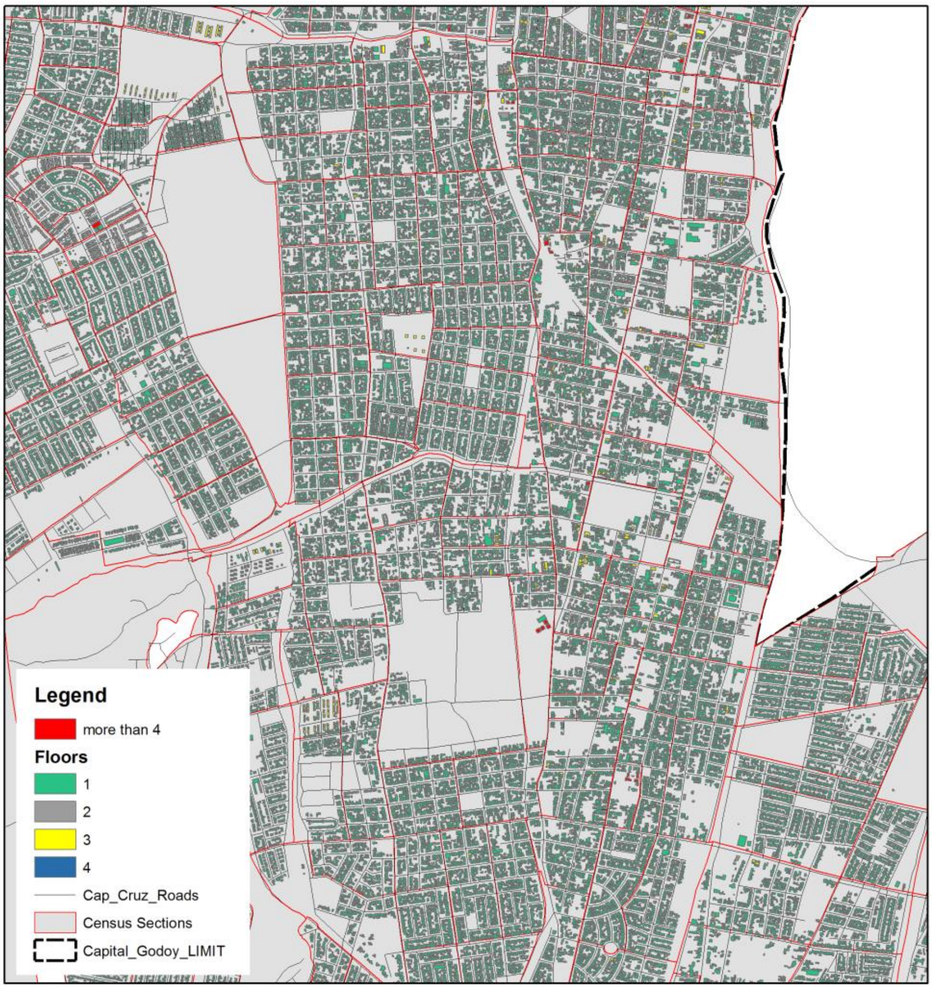


Figure 10. Map of Building's Height of Floor Number. ***Source:*** Author

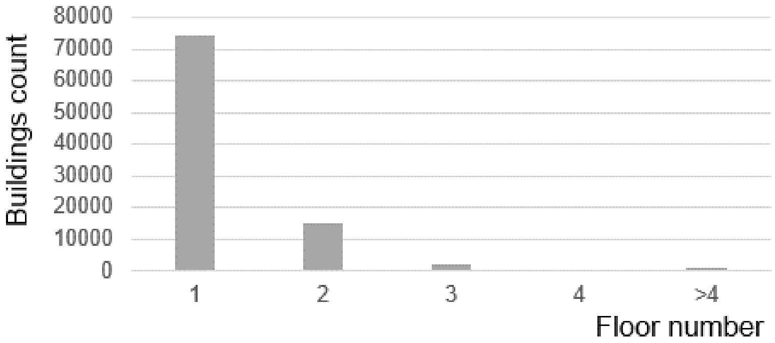


Figure 11. Buildings' Floor Number Count of The Study Area. ***Source:*** Author

4.1.5. Cadastral and DTM

For convenience, cadastral data are utilized to check for the buildings' shapefile, and then the remotely sensed acquired data is filtered to consider eliminating low captured regions from the area considered to be the building, such as garages and canopies attached to the structure.

The DTM was employed with an accuracy of 28m to provide the height of the area, which ranged from +500m to +6530m and reflected the western mountains of Mendoza, as seen below (**Figure 12**).

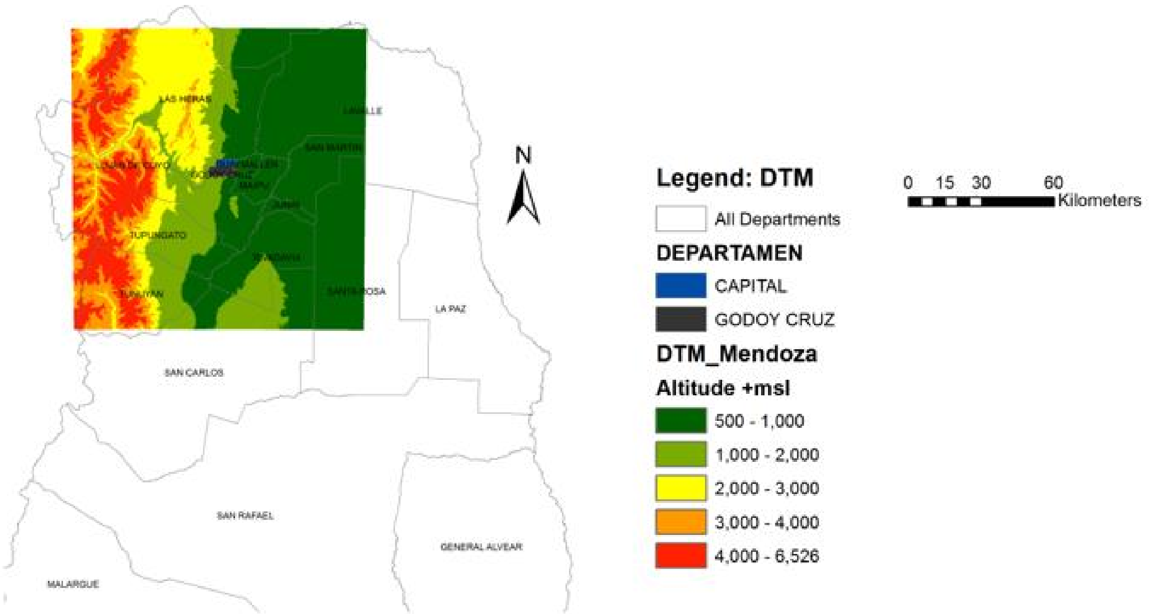


Figure 12. DTM Map of Accuracy 28m for The Study Area. ***Source:*** GIS by Author

4.1.6. Climate Data

The province of Mendoza is located in an area with a large temperature fluctuation due to topography, altitude, and a low rainfall parameter. Summer is considered hot, with an average air temperature exceeding 25 °C, and winter is considered cold and dry, with temperature values below 8 °C. The city's climate is heavily impacted by its arid subtropical location, and the presence of air irrigation canals combined with trees along roadways affects the microclimate, especially during the summer, making the weather bearable even during the warmest hours of the day (Mutani, 2018).

The meteorological analysis looked at the data from four weather stations, which are depicted in the map below, for the reference year of 2016. The weather stations are positioned at various altitudes, as indicated in the graph below (**Figure 13**). The U-shaped temperature distribution shows the seasonal difference, with the summer months being December, January, and February and the winter months being June, July, and August (refer to **Table 5** for the Cooling Degree Days at two of the four weather stations).

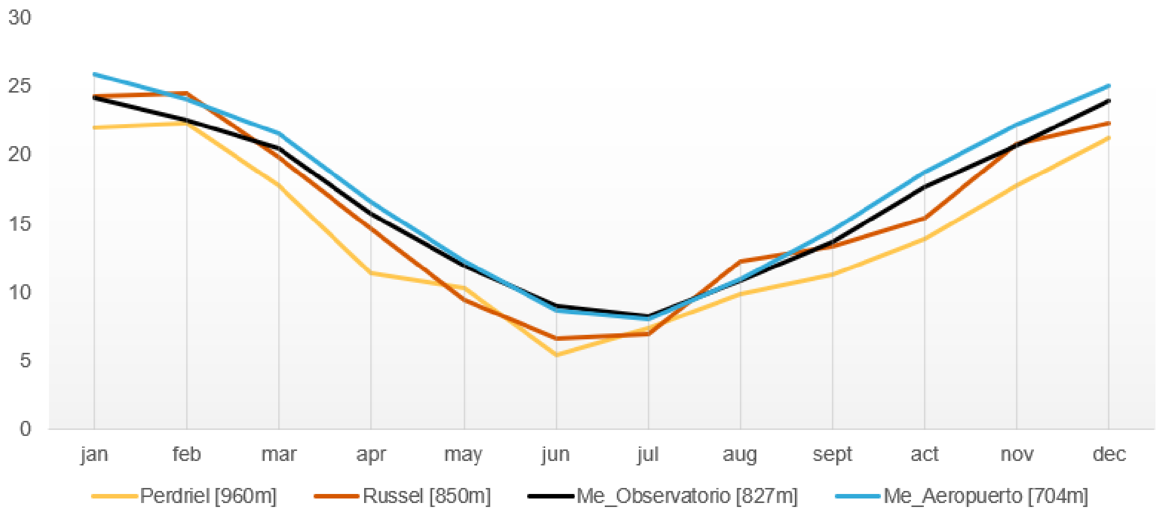


Figure 13. The Recorded Average Temperature of the 4 Stations of Study. ***Source:*** WMO Site

Estación	NRO	NroOACI	tos Climáticos		Unidad	ANUAL	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SET	OCT	NOV	DIC
Mendoza (Aeropuerto)	87418	SAME	a enfriamie	Gdia 23	ºC	215.4	74.6	43.2	10.8	0	0	0	0	0	0	3.4	23.9	59.5
Mendoza (Observatorio)	87420	SAMK	a enfriamie	Gdia 23	ºC	138.3	51.5	30.2	10.3	0	0	0	0	0	0	2.4	11	32.9

Table 5. Cooling Degree Days (CDD) of Observetario and Aeropuerto Weather Stations. ***Source:*** CONICET

The data collected from the four weather stations should then be normalized based on the altitude difference and the average daylight hour per month to check data consistency (**Figure 14**). The normalization procedure is detailed in detail throughout the preprocessing step.

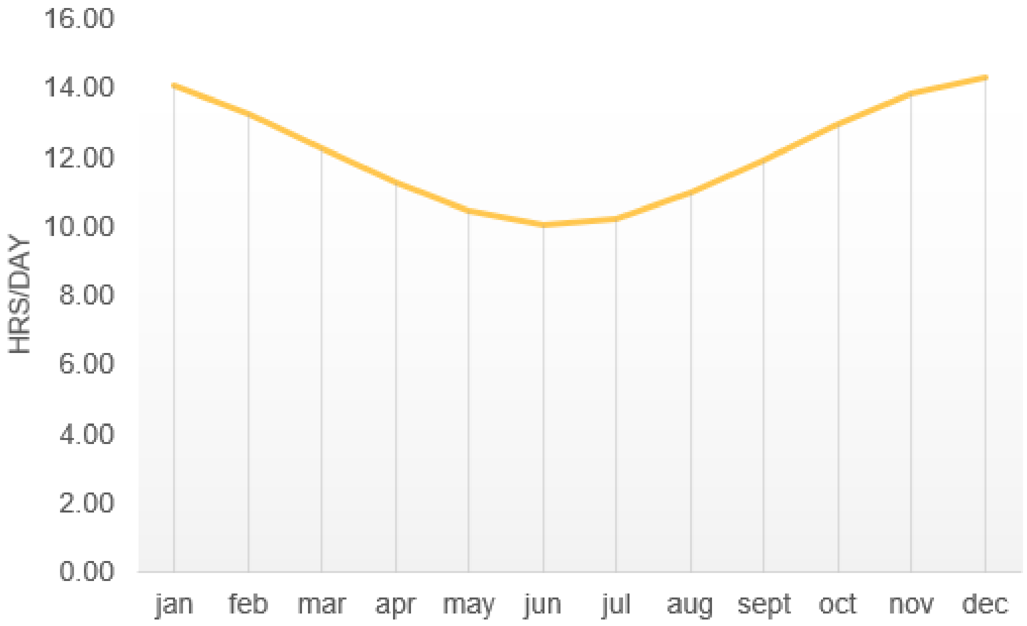


Figure 14. Average Daylight Hours Per Month Of 2016. ***Source:*** Author

4.2. DATA PRE-PROCESSING

The pre-processing phase is critical for the analysis process since it ensures that the data is clean and filtered, as well as that anomalies and outliers are removed. The data is filtered and cleaned in the same order as it was defined. The user consumption clustering procedure was built by starting with the recorded bills of 12 and 6-months for the annual consumption as reported by the EPRE (Provincial Electrical Regulatory Body) and then excluding users with incomplete bills from the annual model. The annual usage of a total of 70,300 consumers is referenced and allocated in the two departments.

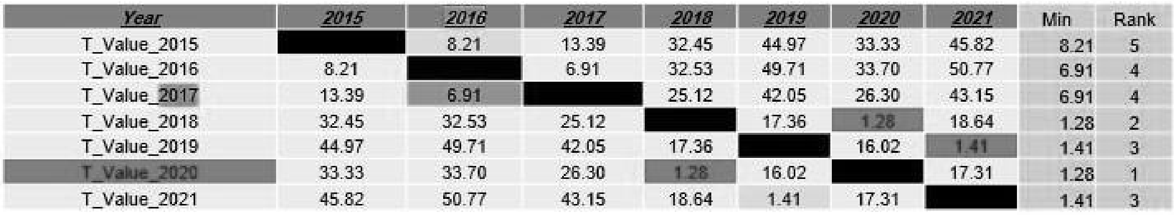
4.2.1. Comparing Datasets

The process of comparing datasets is done to consider the training and corresponding validating data, where the use of student t.test is done by applying the equation of unequal variance test (**Equation 1**), and the result matrix shows the set of lowest rank that should be selected, in addition to considering that the reference year of census data is 2010, and excluding 2020 for the emerging or COVID-19 to avoid the impact of lockdown.

$$\text{T-value} = \frac{mean1 - mean2}{\sqrt{\left(\frac{var1}{n1} + \frac{var2}{n2}\right)}}$$

Equation 1. t.test T-Value of Unequal Variance Sample.

The results were then compared for the selection process, with the year 2016 chosen as a testing reference year and 2017 for validation in the case of temporal analysis of distinct time factor data, such as the average monthly temperature effect (**Table 6**).



Year	2015	2016	2017	2018	2019	2020	2021	Min	Rank
T_Value_2015		8.21	13.39	32.45	44.97	33.33	45.82	8.21	5
T_Value_2016	8.21		6.91	32.53	49.71	33.70	50.77	6.91	4
T_Value_2017	13.39	6.91		25.12	42.05	26.30	43.15	6.91	4
T_Value_2018	32.45	32.53	25.12		17.36	1.28	18.64	1.28	2
T_Value_2019	44.97	49.71	42.05	17.36		16.02	1.41	1.41	3
T_Value_2020	33.33	33.70	26.30	1.28	16.02		17.31	1.28	1
T_Value_2021	45.82	50.77	43.15	18.64	1.41	17.31		1.41	3

Table 6. T-Value Matrix for the Purpose of Selecting Training Dataset Reference Year. ***Source:*** Author

4.2.2. Data Normality Check

The distribution of data influences the selection of appropriate statistical methods based on the kind of distribution, where the normality check procedure is critical to determining whether the methods used for the data set are parametric or non-parametric, considering the violation of normality. Different methodologies, both graphically and numerically, propose the kind of distribution of datasets; initially, a frequency graph of the histogram (**Figure 15**), a qq-plot graph (**Figure 17**), and a box-plot illustrate graphically the distribution of data (**Figure 16**), followed by various quantitative tests.

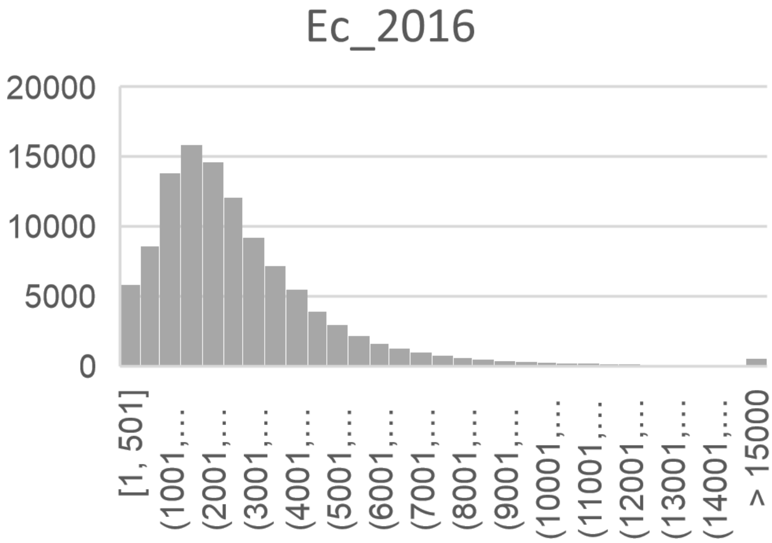


Figure 15. Histogram Of Ec of Reference Year 2016. ***Source:*** Author

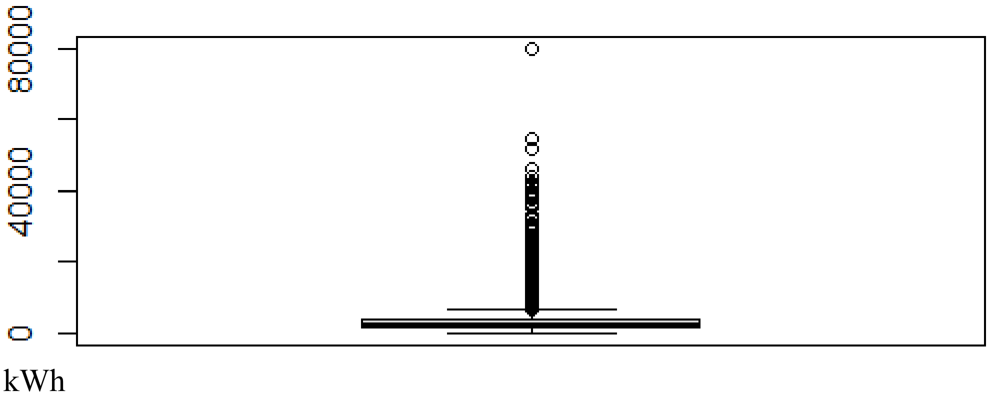


Figure 16. Box-Plot Showing the Median and Outliers of the Distribution. ***Source:*** Author

When the data has a skewed distribution that resembles the gamma or chi-squared distribution, an R-studio software script is used to test the Cullen and Frey graph for the proper distribution curve, as illustrated (**Figure 17** and **Figure 18**).

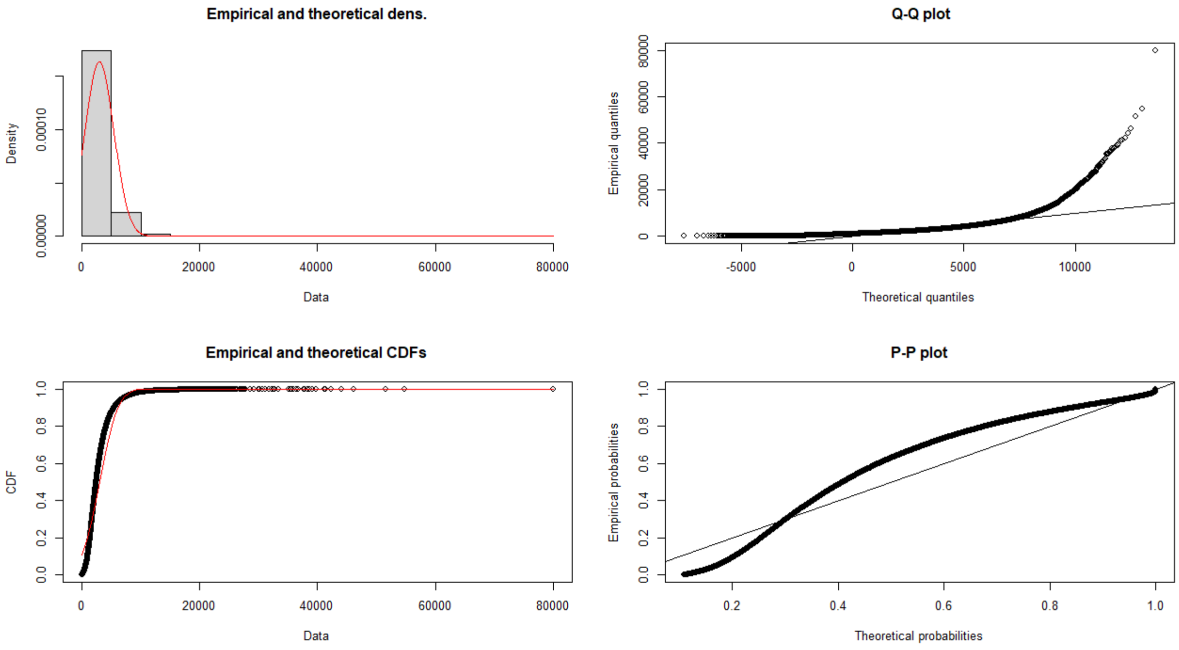


Figure 17. Normal Distribution Check Using R-studio, qq-plot, pp-plot of Data. ***Source:*** Author

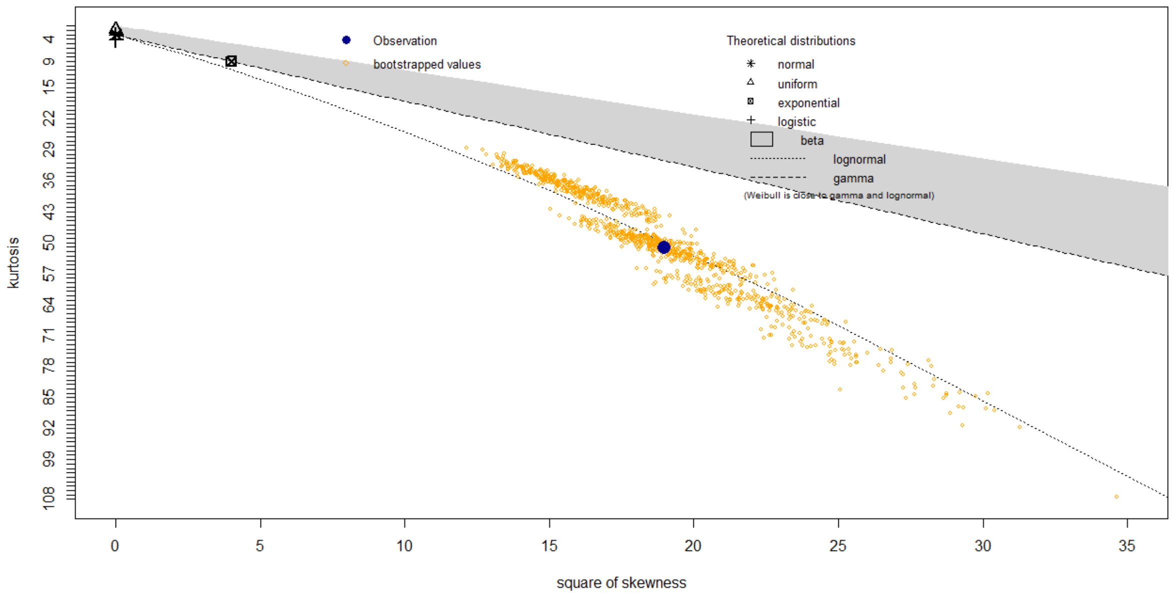


Figure 18. Cullen And Frey Graph of Data Distribution Using R-Studio. ***Source:*** Author

The output depicts the distribution of electricity consumption data as a lognormal type, which must be corrected by using the log of real data (**Figure 19**).

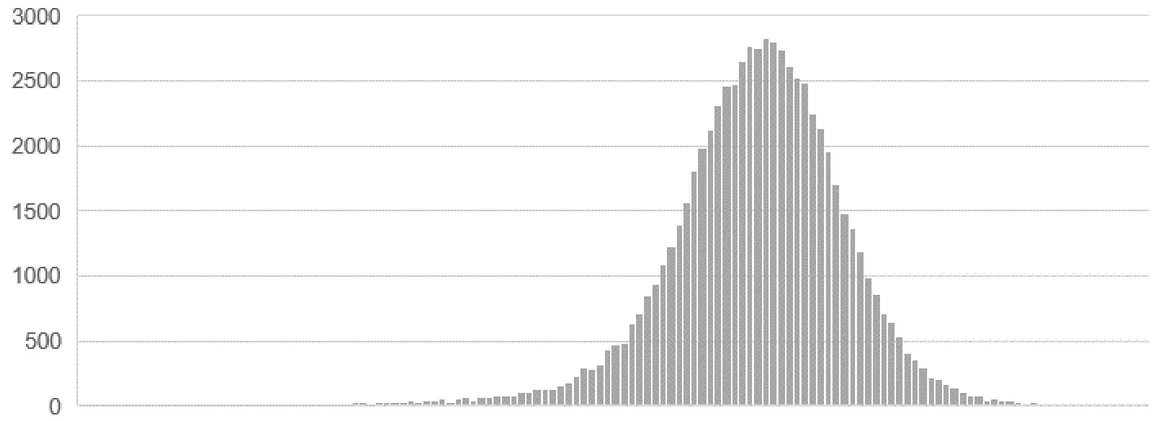
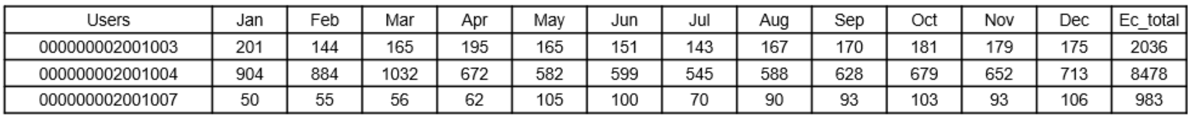


Figure 19. Log-Normal Distribution of Real Ec Data 2016. ***Source:*** Author

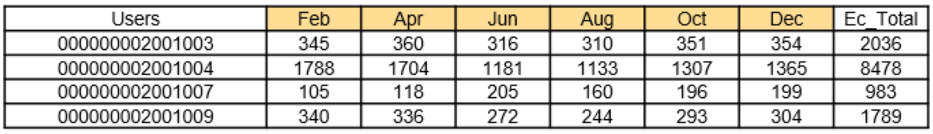
4.2.3. Ec Bills Reorganizing

The disparity in 12-month and 6-month annual consumption discovered during data definition and inquiry indicates the need for reorganising billing data to account for aggregation (**Table 7**). The 12-month Godoy Cruz invoices are re-aggregated to meet the bi-monthly Capital type (**Table 8**).



Users	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ec_total
000000002001003	201	144	165	195	165	151	143	167	170	181	179	175	2036
000000002001004	904	884	1032	672	582	599	545	588	628	679	652	713	8478
000000002001007	50	55	56	62	105	100	70	90	93	103	93	106	983

Table 7. Sample of Godoy Cruz's 12-Month Record. ***Source:*** Author



Users	Feb	Apr	Jun	Aug	Oct	Dec	Ec_Total
000000002001003	345	360	316	310	351	354	2036
000000002001004	1788	1704	1181	1133	1307	1365	8478
000000002001007	105	118	205	160	196	199	983
000000002001009	340	336	272	244	293	304	1789

Table 8. Reorganizing 12-Month Records to Bimonthly. ***Source:*** Author

The bimonthly invoices are then reorganized according to the even and odd months sequence, with additional attributes added and taken into account for the reference sequence (**Figure 20**).

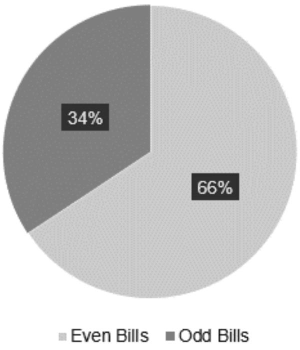
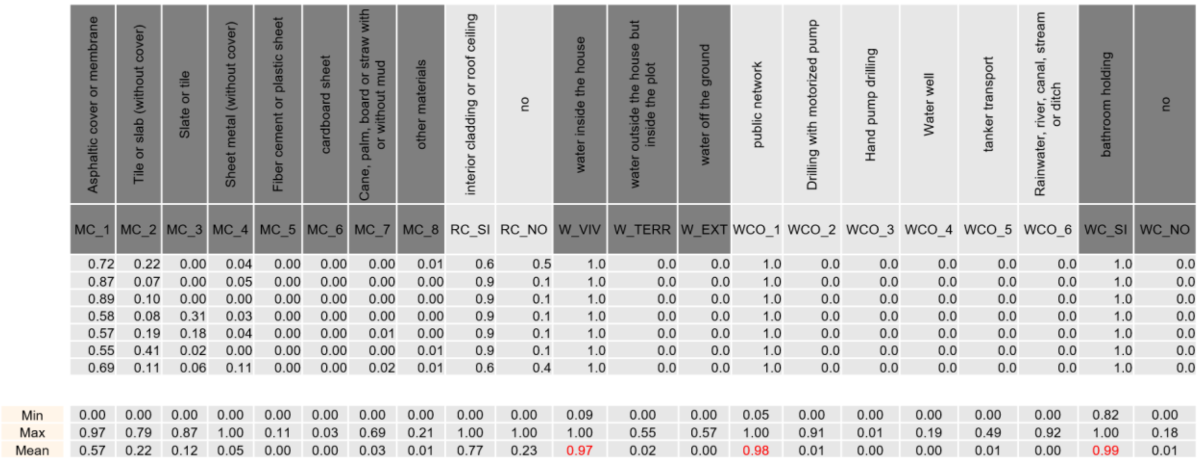


Figure 20. Even/Odd composition of Bills in Capital and Godoy Cruz. ***Source:*** Author

4.2.4. Census Attributes Normalization

The census data of characteristics recorded on the census section shapefile includes the number of families, households, and persons for each category, with the total sum of each attribute category representing the number of families or homes per census zone. To prepare the characteristics for correlation analysis, percentage normalization is performed by dividing the average of each attribute in the category section by the entire amount (**Figure 21**).



	Asphaltic cover or membrane	Tile or slab (without cover)	Slate or tile	Sheet metal (without cover)	Fiber cement or plastic sheet	cardboard sheet	Cane, palm, board or straw with or without mud	other materials	interior cladding or roof ceiling	no	water inside the house	water outside the house but inside the plot	water off the ground	public network	Drilling with motorized pump	Hand pump drilling	Water well	tanker transport	Rainwater, river, canal, stream or ditch	bathroom holding	no
	MC_1	MC_2	MC_3	MC_4	MC_5	MC_6	MC_7	MC_8	RC_SI	RC_NO	W_VIV	W_TERR	W_EXT	WCO_1	WCO_2	WCO_3	WCO_4	WCO_5	WCO_6	WC_SI	WC_NO
	0.72	0.22	0.00	0.04	0.00	0.00	0.00	0.01	0.6	0.5	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	0.87	0.07	0.00	0.05	0.00	0.00	0.00	0.00	0.9	0.1	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	0.89	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.9	0.1	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	0.58	0.08	0.31	0.03	0.00	0.00	0.00	0.00	0.9	0.1	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	0.57	0.19	0.18	0.04	0.00	0.00	0.01	0.00	0.9	0.1	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	0.55	0.41	0.02	0.00	0.00	0.00	0.00	0.01	0.9	0.1	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	0.69	0.11	0.06	0.11	0.00	0.00	0.02	0.01	0.6	0.4	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.82	0.00
Max	0.97	0.79	0.87	1.00	0.11	0.03	0.69	0.21	1.00	1.00	1.00	0.55	0.57	1.00	0.91	0.01	0.19	0.49	0.92	1.00	0.18
Mean	0.57	0.22	0.12	0.05	0.00	0.00	0.03	0.01	0.77	0.23	0.97	0.02	0.00	0.98	0.01	0.00	0.00	0.01	0.00	0.99	0.01

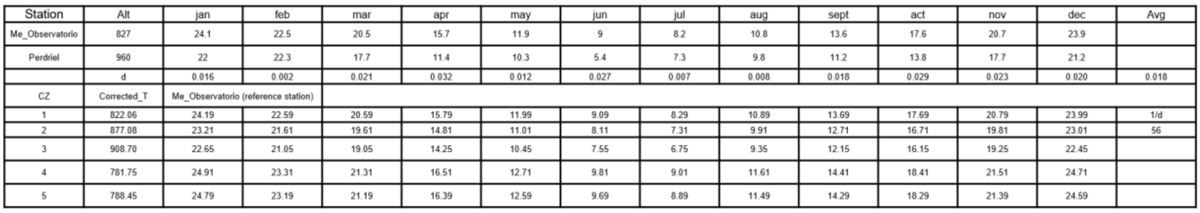
Figure 21. Sample of Percentage Normalized Attributes of CZ After Cleaning Anomalies. ***Source:*** Author

4.2.5. Climate Data Normalization

Using the data provided for the four weather stations, the main Observatorio in Mendoza Capital is chosen as the reference station, while the Perdriel station is used for the equation of distributed temperature by altitude difference (**Equation 2**). The equation was adapted from the March 2016 edition of the Italian standard for Heating and Cooling of Buildings, UNI 10349-1. The procedure was carried out on the Me-Observatorio and Me-Perdriel stations, which share the same features of being located within an urban area and having a close height difference (**Table 9**).

$$\theta_e = \theta_{e,r} - (z - z_r) \times d$$

Equation 2. Equation for Distributing Average Temperature According to Altitude Difference



Station	Alt	jan	feb	mar	apr	may	jun	jul	aug	sept	act	nov	dec	Avg
Me_Observatorio	827	24.1	22.5	20.5	15.7	11.9	9	8.2	10.8	13.6	17.6	20.7	23.9	
Perdriel	960	22	22.3	17.7	11.4	10.3	5.4	7.3	9.8	11.2	13.8	17.7	21.2	
	d	0.016	0.002	0.021	0.032	0.012	0.027	0.007	0.008	0.018	0.029	0.023	0.020	0.018
CZ	Corrected_T	Me_Observatorio (reference station)												
1	822.06	24.19	22.59	20.59	15.79	11.99	9.09	8.29	10.89	13.69	17.69	20.79	23.99	1/d
2	877.08	23.21	21.61	19.61	14.81	11.01	8.11	7.31	9.91	12.71	16.71	19.81	23.01	56
3	908.70	22.65	21.05	19.05	14.25	10.45	7.55	6.75	9.35	12.15	16.15	19.25	22.45	
4	781.75	24.91	23.31	21.31	16.51	12.71	9.81	9.01	11.61	14.41	18.41	21.51	24.71	
5	788.45	24.79	23.19	21.19	16.39	12.59	9.69	8.89	11.49	14.29	18.29	21.39	24.59	

Table 9. Normalized Average Temperature According to Altitude Difference of Census Zones. ***Source:*** Author

4.3. DATA TESTING AND CLUSTERING

Cluster analysis is a method of grouping data using an algorithmic technique with no prior grouping. Unsupervised load profile clustering is extensively used for identifying typical users or buildings based on linked factors (Toussaint & Moodle, 2020). The clustering algorithm used for the consumers is K-means clustering of Euclidian distance in two stages: one for the total consumption as it is without considering any additional characteristics, and the other for the log-consumption and the percentages of variables obtained from the census section level and joined spatially to users inhabiting these sections. The clustering findings are then categorized using a simple user categorization based on energy consumption, which is represented by low and high consumers with an intermediate level.

4.3.1. First Cluster Analysis

Before running the K-means algorithm, the frequency curve of Ec data from the reference year 2016 is checked (**Figure 22**), and the supervised categorization of percentages according to a random bin is performed (**Table 10**).

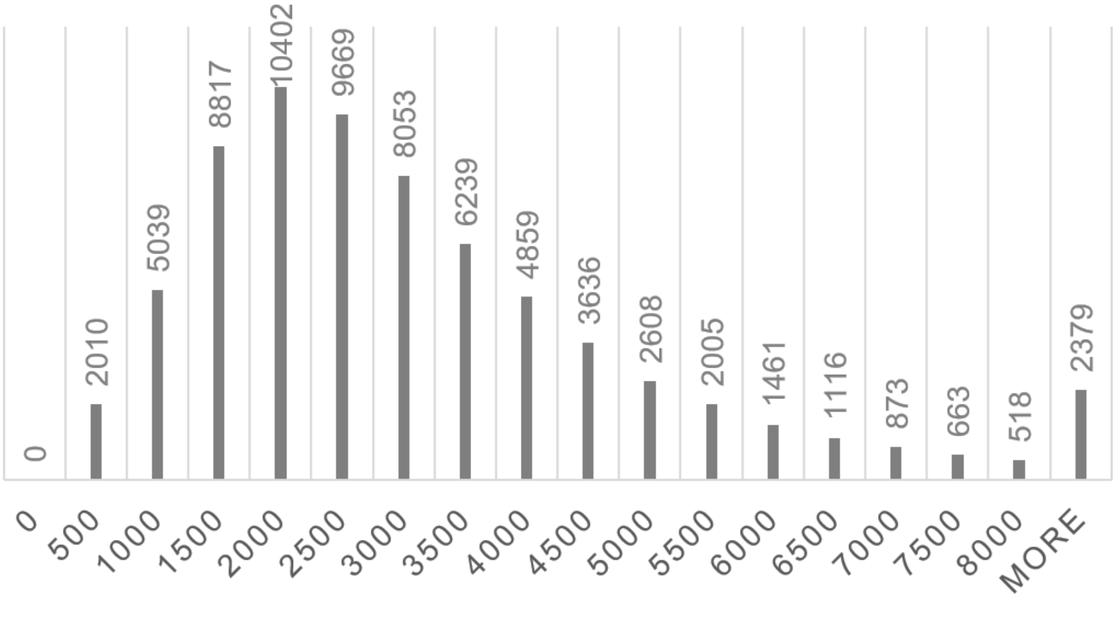
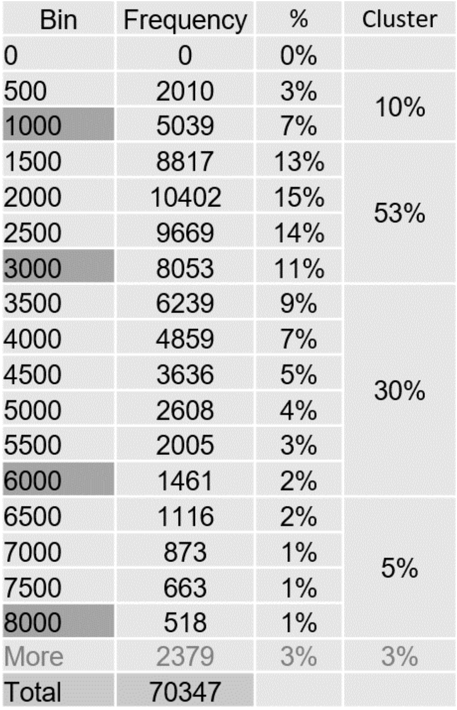


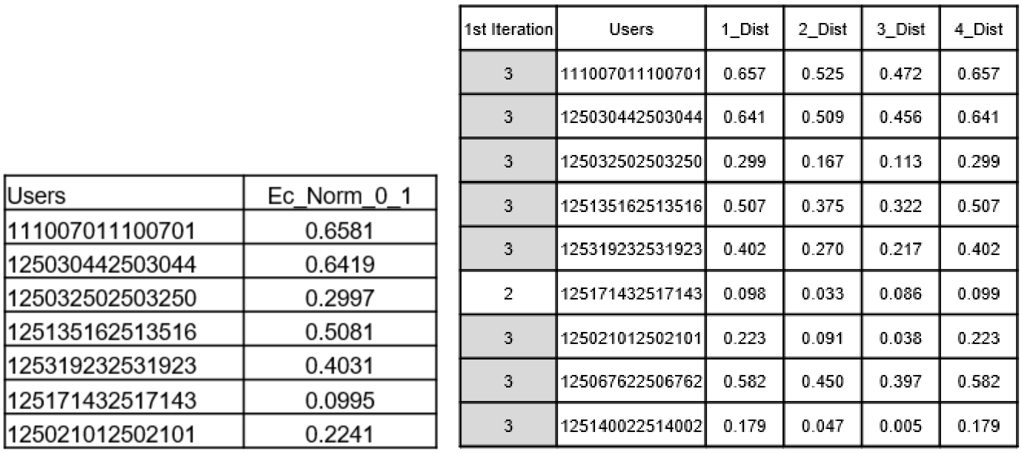
Figure 22. Users Ec Frequency Curve 2016. ***Source:*** Author



Bin	Frequency	%	Cluster
0	0	0%	
500	2010	3%	10%
1000	5039	7%	
1500	8817	13%	53%
2000	10402	15%	
2500	9669	14%	
3000	8053	11%	
3500	6239	9%	30%
4000	4859	7%	
4500	3636	5%	
5000	2608	4%	
5500	2005	3%	
6000	1461	2%	
6500	1116	2%	5%
7000	873	1%	
7500	663	1%	
8000	518	1%	
More	2379	3%	3%
Total	70347		

Table 10. Users Ec Frequency Supervised Random Bin 2016. ***Source:*** Author

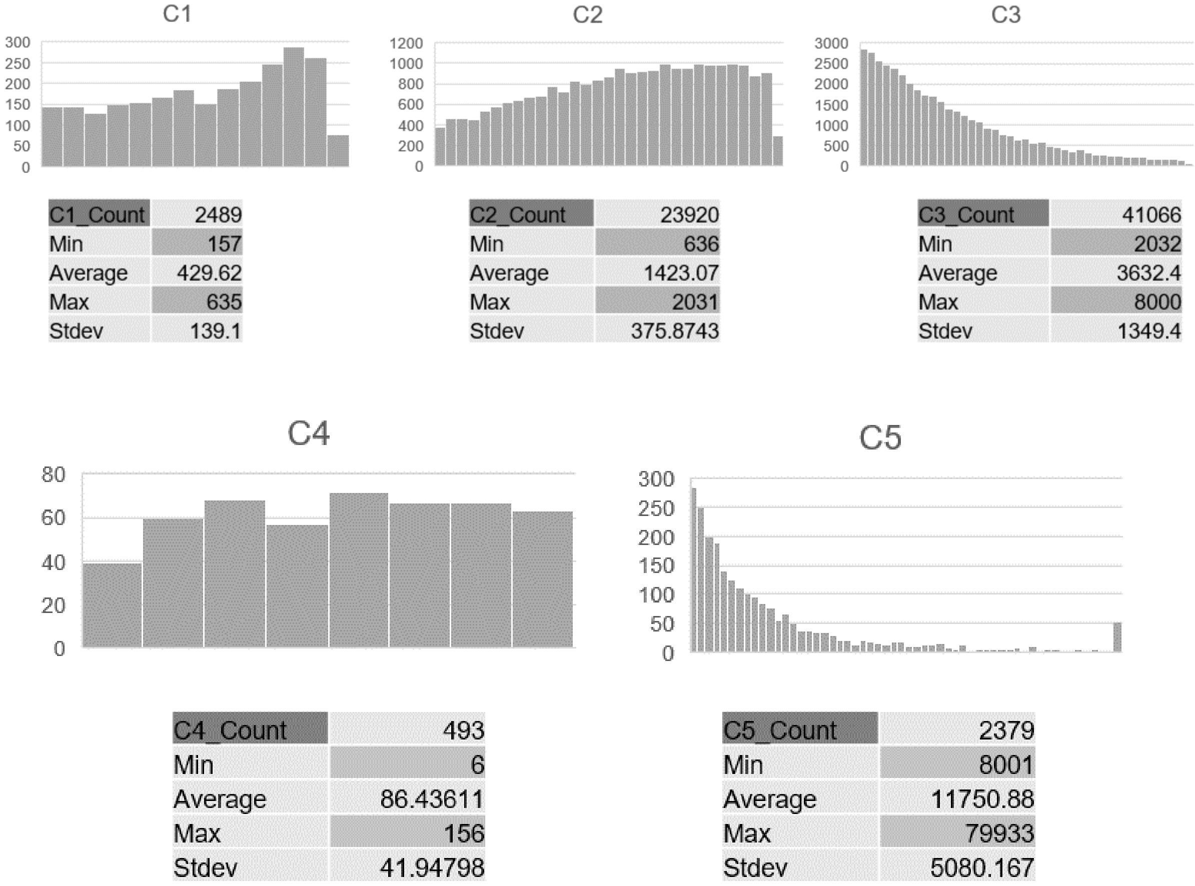
For faster and better computation, the K-means algorithm is performed manually in Excel after normalizing the user's Ec to 0 and 1 normalization procedure (**Table 11** and **Figure 23**).



Users	Ec_Norm_0_1
111007011100701	0.6581
125030442503044	0.6419
125032502503250	0.2997
125135162513516	0.5081
125319232531923	0.4031
125171432517143	0.0995
125021012502101	0.2241

1st Iteration	Users	1_Dist	2_Dist	3_Dist	4_Dist
3	111007011100701	0.657	0.525	0.472	0.657
3	125030442503044	0.641	0.509	0.456	0.641
3	125032502503250	0.299	0.167	0.113	0.299
3	125135162513516	0.507	0.375	0.322	0.507
3	125319232531923	0.402	0.270	0.217	0.402
2	125171432517143	0.098	0.033	0.086	0.099
3	125021012502101	0.223	0.091	0.038	0.223
3	125067622506762	0.582	0.450	0.397	0.582
3	125140022514002	0.179	0.047	0.005	0.179

Table 11. Users Ec (0,1) Normalization and Sample of Algorithm Iteration. ***Source:*** Author



C1_Count	2489
Min	157
Average	429.62
Max	635
Stdev	139.1

C2_Count	23920
Min	636
Average	1423.07
Max	2031
Stdev	375.8743

C3_Count	41066
Min	2032
Average	3632.4
Max	8000
Stdev	1349.4

C4_Count	493
Min	6
Average	86.43611
Max	156
Stdev	41.94798

C5_Count	2379
Min	8001
Average	11750.88
Max	79933
Stdev	5080.167

Figure 23. Cluster No.1 for the Absolute Ec, K=5. ***Source:*** Author

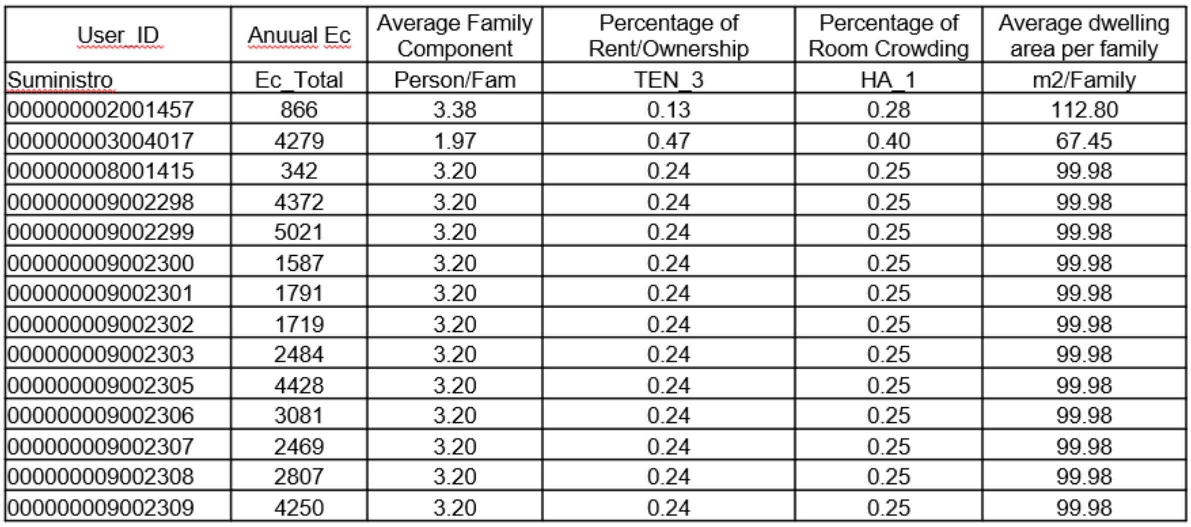
The obtained cluster count is five consumption clusters without respect to any other characteristics of the sample, and a supervised classification was then added to differentiate the clusters based on logical interpretation (**Table 12**).

Consumer Type	User_Count	%	Cluster	Avg_(Jan/Feb)	Avg_(Mar/Apr)	Avg_(May/Jun)	Avg_(Jul/Aug)	Avg_(Sep/Oct)	Avg_(Nov/Dec)
V.Low	493	0.7%	4	15.6	13.8	14.4	14.5	13.8	14.2
Low	2,489	3.5%	1	83.6	72.8	68.2	69.3	64.2	71.8
Medium-Low	23,920	34.0%	2	280.3	251.9	229.5	238.7	207.8	214.9
Medium	41,066	58.4%	3	701.1	611.5	610.9	676.6	526.9	505.4
High	2,379	3.4%	5	2,120.6	1,952.4	2,000.8	2,258.3	1,766.6	1,652.2

Table 12. Classified Users According to The Results of Clustering No.1. ***Source:*** Author

4.3.2. Second Cluster Analysis

The clustering algorithm was used with R-studio software and a library script, and the input data was the normalized Ec of average consumption per person for each census section, as well as the corresponding average normalized socio-demographic characteristics of relative correlation of significance. Users in each census zone are spatially linked to their respective parts, and the census characteristics are averaged accordingly (**Table 13**). The approach required anticipating the optimal K of clusters (**Figure 24**), which is accomplished by WSS analysis, and the resulting clusters are two distinct users (**Table 14**).



User_ID	Anual Ec	Average Family Component	Percentage of Rent/Ownership	Percentage of Room Crowding	Average dwelling area per family
Suministro	Ec_Total	Person/Fam	TEN_3	HA_1	m2/Family
000000002001457	866	3.38	0.13	0.28	112.80
000000003004017	4279	1.97	0.47	0.40	67.45
000000008001415	342	3.20	0.24	0.25	99.98
000000009002298	4372	3.20	0.24	0.25	99.98
000000009002299	5021	3.20	0.24	0.25	99.98
000000009002300	1587	3.20	0.24	0.25	99.98
000000009002301	1791	3.20	0.24	0.25	99.98
000000009002302	1719	3.20	0.24	0.25	99.98
000000009002303	2484	3.20	0.24	0.25	99.98
000000009002305	4428	3.20	0.24	0.25	99.98
000000009002306	3081	3.20	0.24	0.25	99.98
000000009002307	2469	3.20	0.24	0.25	99.98
000000009002308	2807	3.20	0.24	0.25	99.98
000000009002309	4250	3.20	0.24	0.25	99.98

Table 13. Users Aggregated Per Census Section and Assigned Corresponding Average Characteristics. ***Source:*** Author

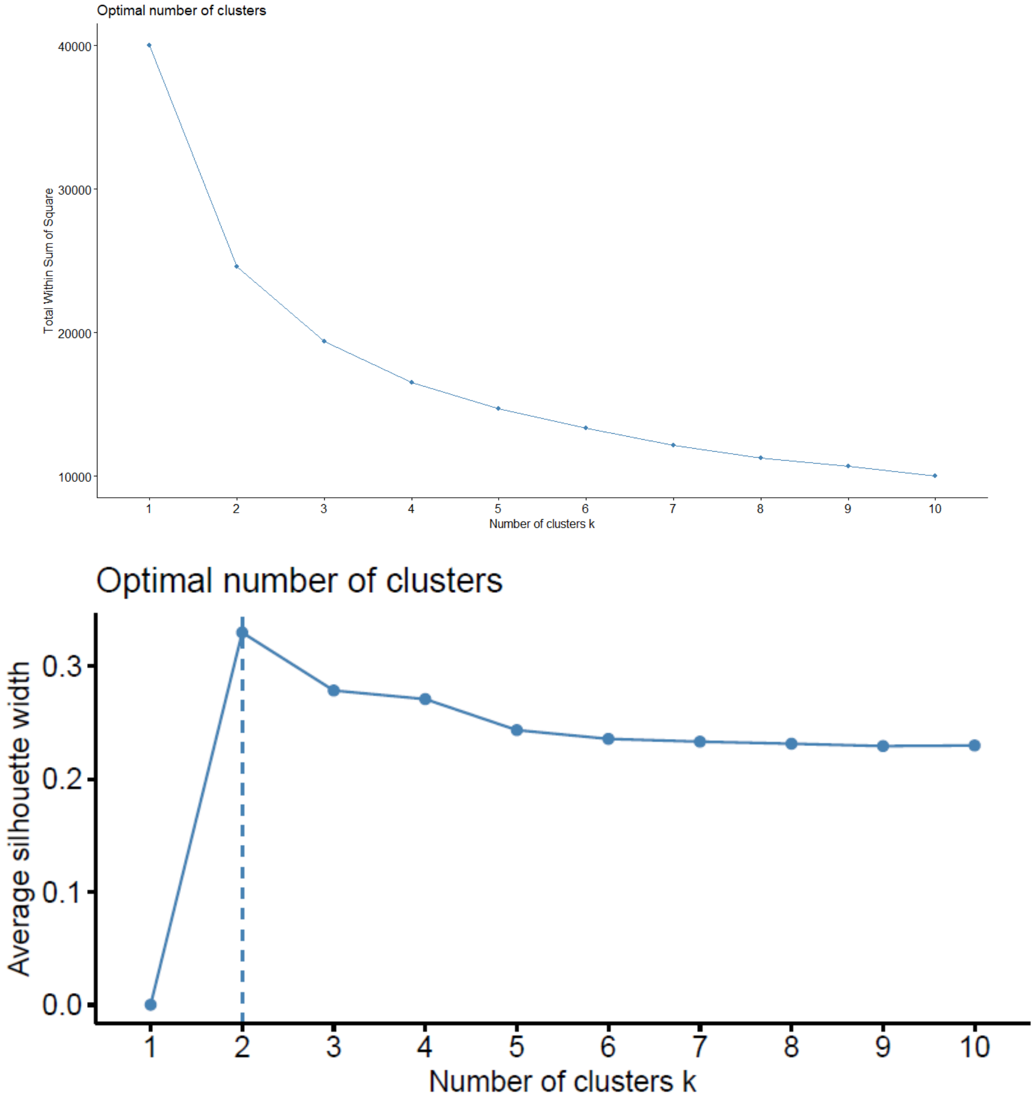
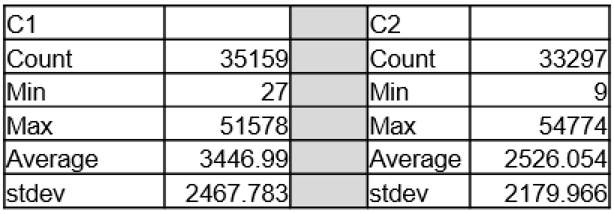


Figure 24. Optimal K Cluster for The Input Data. ***Source:*** Author

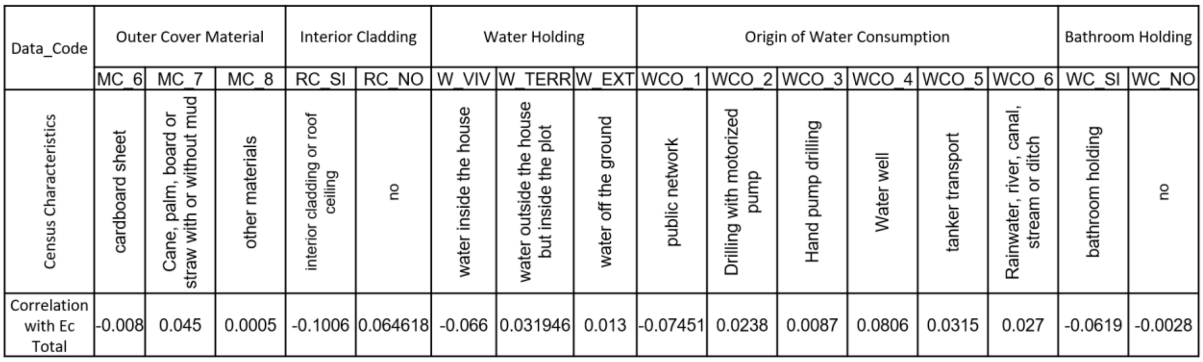


C1			C2	
Count	35159		Count	33297
Min	27		Min	9
Max	51578		Max	54774
Average	3446.99		Average	2526.054
stdev	2467.783		stdev	2179.966

Table 14. Result Of Clustering No 2. ***Source:*** Author

4.4. VARIABLES CORRELATION MATRIX

The correlation analysis is performed to identify the highly correlated independent variables that are assigned to users from census section attributes, where the records are normalized and then correlated to the total annual consumption per user, and the results revealed a poor correlation with all socio-demographic factors regarding the average user consumption per census section (**Table 15**).

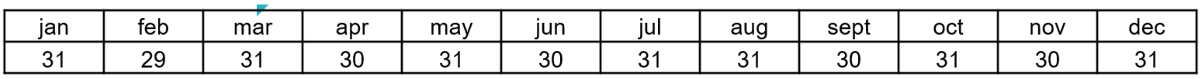


Data_Code	Outer Cover Material			Interior Cladding		Water Holding			Origin of Water Consumption						Bathroom Holding	
	MC_6	MC_7	MC_8	RC_SI	RC_NO	W_VIV	W_TERR	W_EXT	WCO_1	WCO_2	WCO_3	WCO_4	WCO_5	WCO_6	WC_SI	WC_NO
Census Characteristics	cardboard sheet	Cane, palm, board or straw with or without mud	other materials	interior cladding or roof ceiling	no	water inside the house	water outside the house but inside the plot	water off the ground	public network	Drilling with motorized pump	Hand pump drilling	Water well	tanker transport	Rainwater, river, canal, stream or ditch	bathroom holding	no
Correlation with Ec Total	-0.008	0.045	0.0005	-0.1006	0.064618	-0.066	0.031946	0.013	-0.07451	0.0238	0.0087	0.0806	0.0315	0.027	-0.0619	-0.0028

Table 15. Sample Correlation matrix between Ec total of average user per census section and the socio-demographic Factors. ***Source:*** Author

4.4.1. Correlation with Average Temperature

The average temperature is added to the dataset based on the change in monthly temperature, with a pre-step of averaging the temperature on a biweekly basis to correlate with billing consumption. The monthly day count is taken into account for each month in 2016, and the average daily temperature is determined using the equation (**Table 16**).



jan	feb	mar	apr	may	jun	jul	aug	sept	oct	nov	dec
31	29	31	30	31	30	31	31	30	31	30	31

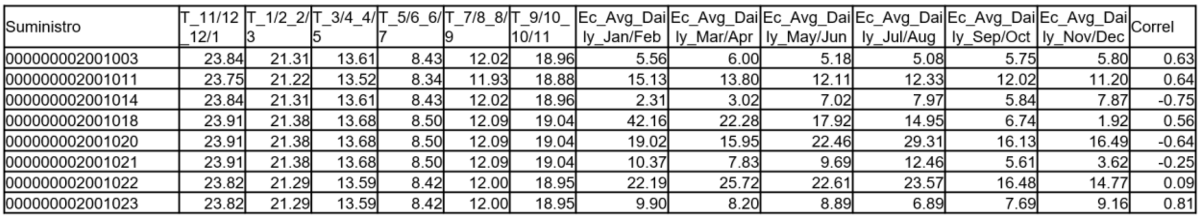
Table 16. Monthly Days Count of Reference Year 2016**.** ***Source:*** Author

Ex. Nov = 20.7 C° & Dec = 23.9 C°, The average T = (20.7*30+23.9*31)/(30+31) = 22.32 C°

Similarly, the average daily electrical consumption is calculated for considering the correlation consistency (**Table 17**).

Ex. Bimonthly bill of Jan = consumption of November + December, if bill of Jan = 345 kWh.

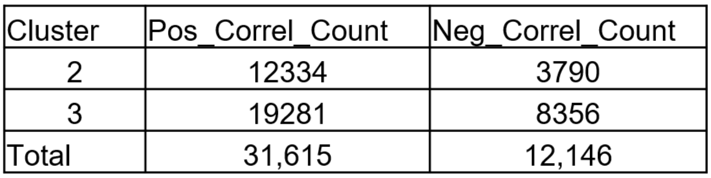
Then Avg daily consumption = Ec(Daily) = 345 / (30+31) = 5.65 kWh/Day.



Suministro	T_11/12_12/1	T_1/2_2/3	T_3/4_4/5	T_5/6_6/7	T_7/8_8/9	T_9/10_10/11	Ec_Avg_Daily_Jan/Feb	Ec_Avg_Daily_Mar/Apr	Ec_Avg_Daily_May/Jun	Ec_Avg_Daily_Jul/Aug	Ec_Avg_Daily_Sep/Oct	Ec_Avg_Daily_Nov/Dec	Correl
000000002001003	23.84	21.31	13.61	8.43	12.02	18.96	5.56	6.00	5.18	5.08	5.75	5.80	0.63
000000002001011	23.75	21.22	13.52	8.34	11.93	18.88	15.13	13.80	12.11	12.33	12.02	11.20	0.64
000000002001014	23.84	21.31	13.61	8.43	12.02	18.96	2.31	3.02	7.02	7.97	5.84	7.87	-0.75
000000002001018	23.91	21.38	13.68	8.50	12.09	19.04	42.16	22.28	17.92	14.95	6.74	1.92	0.56
000000002001020	23.91	21.38	13.68	8.50	12.09	19.04	19.02	15.95	22.46	29.31	16.13	16.49	-0.64
000000002001021	23.91	21.38	13.68	8.50	12.09	19.04	10.37	7.83	9.69	12.46	5.61	3.62	-0.25
000000002001022	23.82	21.29	13.59	8.42	12.00	18.95	22.19	25.72	22.61	23.57	16.48	14.77	0.09
000000002001023	23.82	21.29	13.59	8.42	12.00	18.95	9.90	8.20	8.89	6.89	7.69	9.16	0.81

Table 17. Average Daily Temperature Correlation to Average Daily User Consumption. ***Source:*** Author

The results of correlation then classified into positive of more than 0.4, and negative of less than -0.4 correlation coefficient according to cluster number. The result shows about 50% of users are correlated to seasonal consumption (**Table 18**).



Cluster	Pos_Correl_Count	Neg_Correl_Count
2	12334	3790
3	19281	8356
Total	31,615	12,146

Table 18. Positive and Negative Correlation of C2 and C3. ***Source:*** Author

4.4.2. Correlation by Average Ec per Person

The process of reprocessing is done to further try another dependent variable of Ec/person/census zone, the variable that is calculated according to the average consumption of users in each census zone divided by the average person per family to give the consumption per person per census zone. Higher correlation coefficient appeared regarding three variables, the family component ratio, dwelling ownership ratio, and room crowding ratio.

4.5. DATA MODELLING AND VALIDATING

4.5.1. Linear Regression for Average Temperature Test

Firstly, the data of bimonthly consumption is used against the change in temperature to test the significance of regression analysis according to this assumption of seasonal effect, considering the positively correlated users in the main two clusters of first clustering analysis, C2 and C3. The report of regression showed that, in case of C3 positive correlated users, the R2 is only (0.16) which does not validate most of the records (**Figure 25**).

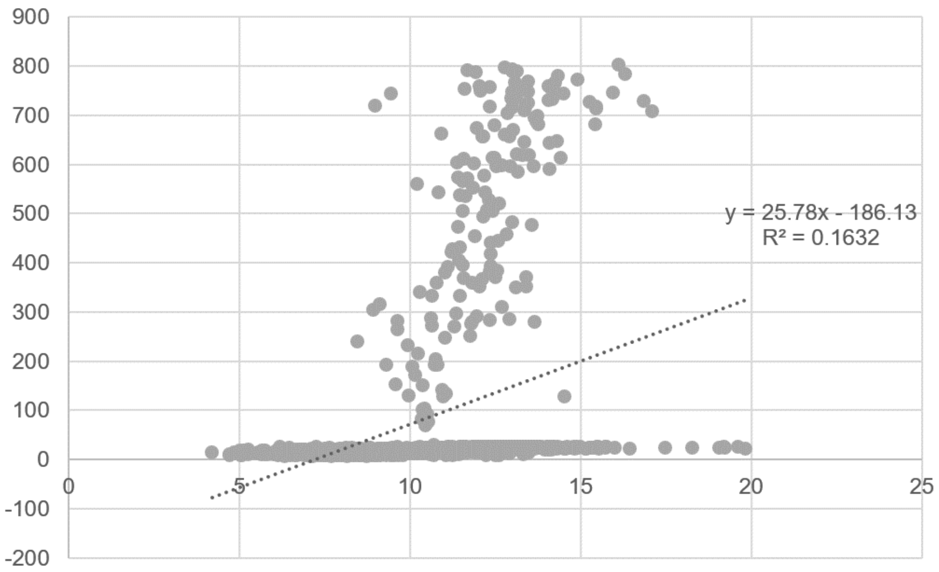


Figure 25. Scattered Plot for The Average Daily Temperature (X-Axis) And the Aggregated Average Daily Consumption Per Census Section (Y-Axis). ***Source:*** Author

4.5.2. Multiple Linear Regression for Selected Variables

The correlation found between average Ec/person/census with the HA1, TEN3, and Person/Family (**Table 19** and **Figure 26**).

Selected Variables	TEN3	HA1	Person/Family
C2016_Ec/Person	0.59	0.60	-0.70

Table 19, Correlation coefficient of selected variables. ***Source:*** Author

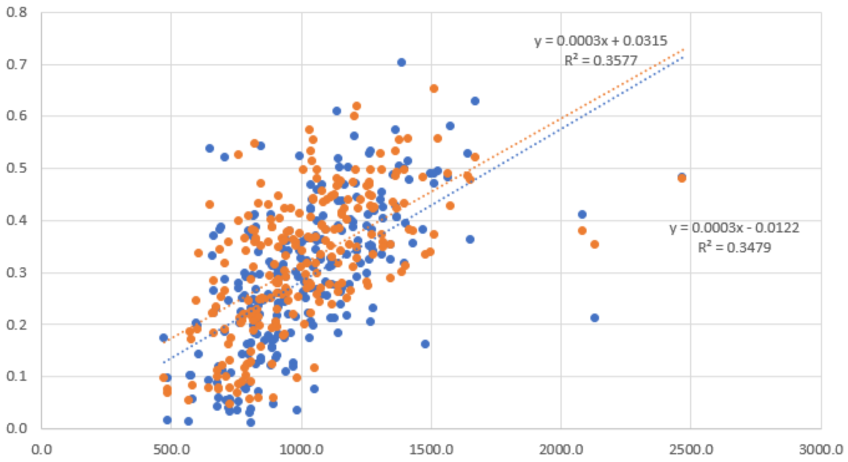
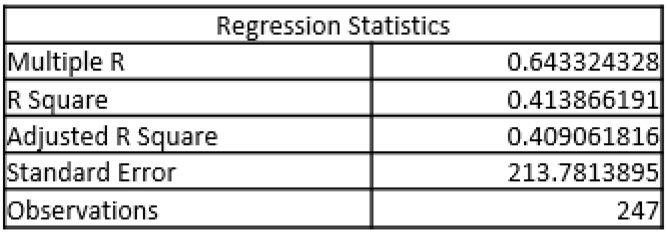


Figure 26, Scattered plot for average Ec/person and HA1(blue) vs TEN3(orange). ***Source:*** Author

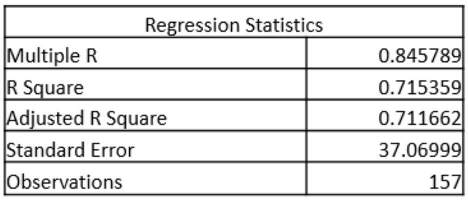
The multiple regression done to C2, and C3, showed there is a significancy of only two variables in each cluster set. For the case of C3 with considering all the Census sections of Capital and Godoy Cruz.



Regression Statistics	
Multiple R	0.643324328
R Square	0.413866191
Adjusted R Square	0.409061816
Standard Error	213.7813895
Observations	247

Table 20, MLR of C3 for the two departments. ***Source:*** Author

Then the dataset is separated to consider the two departments as one for the testing and the other for validating the coefficients of MLR.

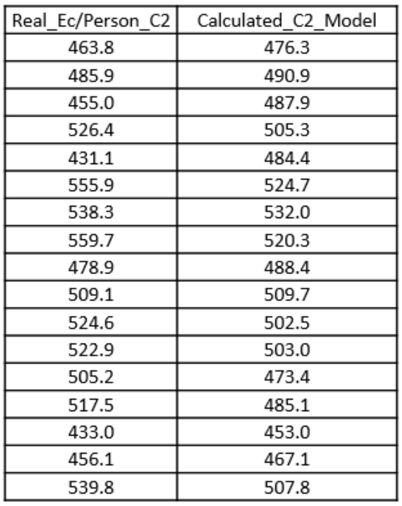


Regression Statistics	
Multiple R	0.845789
R Square	0.715359
Adjusted R Square	0.711662
Standard Error	37.06999
Observations	157

Table 21, R2 of Godoy Cruz C2 with Person/Family and Room Crowding Variables. ***Source:*** Author

Y = 157.62x(HA_1) – 80.06x(Person/Family) + 676.7 (C2 Prediction Eqaution)

The equation of MLR coefficient resulted is then used for the testing of Capital census zones (**Table 22**).



Real_Ec/Person_C2	Calculated_C2_Model
463.8	476.3
485.9	490.9
455.0	487.9
526.4	505.3
431.1	484.4
555.9	524.7
538.3	532.0
559.7	520.3
478.9	488.4
509.1	509.7
524.6	502.5
522.9	503.0
505.2	473.4
517.5	485.1
433.0	453.0
456.1	467.1
539.8	507.8

Table 22, Sample of the real vs estimated Ec of Capital department using the equation of C2. ***Source:*** Author

The model gives an average MAPE of 6% for C2, and when repeated for C3, the MAPE is about 7%. With the following equation.

Y = 296.76x(TEN3) – 262.55x(Person/Family) +1932.77 (C3 Prediction Equation)

5. CONCLUSION

5.1. SUMMARY

The annual electrical consumption billing records reveal variation among consumers, who are classified into five groups. Furthermore, there is a billing system mismatch between the two departments, Mendoza Capital and Godoy Cruz, with 12-month and 6-bimonth billing with many bills missing, particularly in the Godoy Cruz department. The unaccounted-for bills are likely to pertain to the temporary usage of a dwelling, which primarily alludes to active rental activities and/or temporary housing. However, most billing data and annual aggregates in the Mendoza Capital department provide more consistent knowledge for the area's users.

In 2016, approximately 70,300 users were chosen from a total of 113,000 recorded bills for the modelling phase, where the data is divided into two groups: one for aggregated bimonthly-related processing that considers seasonal temperature effects, and the same data aggregated on an annual basis to share in the modelling that considers socio-demographic variables on a census section scale. In the Multiple Linear Regression model of socio-demographic factors, the findings of the two models demonstrate a substantial association between the dwelling of rental activity and the room crowding ratio for each census sector. The consumption is highly associated with the change in mean temperature in the linear Regression model for seasonal change.

5.2. CONCLUSION

Analysing electricity consumption data using LR and MLR models according to the type of variables that share in the model, as the proposed model gives evidence for the impact of rental activity with home crowding that shows how electrical consumption is related to the ownership of dwelling and low room crowding, when related in terms of annual consumption per person, and on the other hand, for the LR model, the effect of seasonal temperature change on about 60% of users that mainly use cooling in summer times, where the change of mean temperature has an effect on seasonal consumption of electricity.

5.3. RECOMMENDATIONS

This study can be used by the Mendoza municipality, particularly the planning and infrastructure departments, to better estimate and plan future extensions and/or renovation processes to efficiently predict resources needed for the future demand for electricity supply for the housing sector and residential projects. Demand forecasting is critical for network architecture and future energy output. Furthermore, policymakers can benefit from the findings by establishing rules that regulate and better moderate electricity consumption, either by using the power of tariffs and subsidies and better targeting prices to users based on dwelling activity or by offering incentives for self-production of energy through solar power and/or compensating consumption by encouraging the use of more energy-efficient appliances.

5.4. FUTURE WORKS

The better the input data, the more accurate the prediction outcomes, with little prediction error. Right now, I would advocate undertaking a field study of a random sample of the population and households to gather more information about direct electricity use in terms of devices and appliances. The study of load profiles, combined with accurate metering for home appliances at daily and hourly intervals, provides more accurate modelling for electrical consumption, where the variability of consumption appears throughout the day between day and night times and shows how consumption varies between weekdays and weekends.

The integration of user socioeconomic and home appliance data into billing systems can improve the quality of data input and modelling processes, where database updates can be self-inserted by users on an annual basis, which could be encouraged by electricity supply companies to their clients in the form of incentives or bill instalment plans. The objective is to create a win-win, sustainable procedure for assessing and updating a department-scale database of power consumers with robust and direct data on use. Data that not only directly consumes energy but also provides consumers with an analysis of their socioeconomic status.

6. REFERENCES

Alqasim, A. R. (2022). *Using Regression Analysis for Predicting Energy Consumption in Dubai Police* (Master's thesis). Rochester, NY: Rochester Institute of Technology.

Chuan, L., & Ukil, A. (2015). Modeling and Validation of Electrical Load Profiling in Residential Buildings in Singapore. *IEEE Transactions on Power Systems, 30*(5), 2800–2809. DOI: 10.1109/TPWRS.2014.2367509

Fumo, N. (2014). A Review on the Basics of Building Energy Estimation. *Renewable and Sustainable Energy Reviews, 31*, 53–60. DOI: 10.1016/j.rser.2013.11.040

Fumo, N., & Rafe Biswas, M. A. (2015). Regression Analysis for Prediction of Residential Energy Consumption. *Renewable and Sustainable Energy Reviews, 47,* 332–343. DOI: 10.1016/j.rser.2015.03.035

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z., & Djurovic-Petrovic, M. (2010). A Review of Bottom-Up Building Stock Models for Energy Consumption in the Residential Sector. *Building and Environment, 45*(7), 1683–1697. DOI: 10.1016/j.builde nv.2010.01.021

Lamagna, M., Nastasi, B., Groppi, D., Nezhad, M. M., & Garcia, D. A. (2020). Hourly Energy Profile Determination Technique From Monthly Energy Bills. *Building Simulation, 13*, 1235–1248. DOI: 10.1007/s12273-020-0698-y

Lazzari, F., Mor, G., Cipriano, J., Gabaldon, E., Grillone, B., Chemisana, D., & Solsona, F. (2022). User behaviour models to forecast electricity consumption of residential customers based on smart metering data. *Energy Reports, 8*, 3680–3691. DOI: 10.1016/j.egyr.2022.02.260

Luo, X., Hong, T., Chen, Y., & Piette, M. A. (2017). Electric Load Shape Benchmarking for Small- And Medium-Sized Commercial Buildings. *Applied Energy, 204*(15), 715–725. DOI: 10.1016/j.apenergy.2017.07.108

Mutani, G., Fontanive, M., & Arboit, M. E. (2018). Energy use Modelling for Residential Buildings in the Metropolitan Area of Gran Mendoza – AR. *TECNICA ITALIANA-Italian Journal of Engineering Science, 61+1*(2), 74–82. DOI: 10.18280/ti-ijes.620204

Mutani, G., Fontana, R., & Barreto, A. (2019). "Spatial energy modelling for the Metropolitan City of Rome", 2021 IEEE 4th International Conference and Workshop Óbuda on Electrical and Power Engineering (CANDO-EPE), pp.43-48

Reta, R. (2012). *Plan Estratégico De Desarrollo Energético Para La Provincia De Mendoza* [Strategic Energy Development Plan for the Province of Mendoza]. Informe Final [online]. Retrieved from http://biblioteca.cfi.org.ar/wp-content/uploads/sites/2/2012/01/49233.pdf

Subba, A. B., Chodon, P., Drukpa, T., Jayachandran, V., & Lhendup, T. (2015). *Residential Electricity Use Load Profile Using Monthly Electricity Consumption.* Conference: International Seminar on Renewable Energy and Sustainable Development.

Swan, L. G. & Ugursal, V. I. (2009). Modeling of End-Use Energy Consumption in the Residential Sector: A Review of Modeling Techniques. *Renewable and Sustainable Energy Reviews, 13*(8), 1819–1835. DOI: 10.1016/j.rser.2008.09.033

Toussaint, W. & Moodle, D. (2020). Clustering Residential Electricity Consumption Data to Create Archetypes that Capture Household Behaviour in South Africa. *SACJ,* 32(2). DOI: 10.18489/sacj.v32i2.845

Zakovorotnyi, A., & Seerig, A. (2017). Building Energy Data Analysis by Clustering Measured Daily Profiles. *Energy Procedia, 122,* 583–588. DOI: 10.1016/j.egypro.2017.07.353