ECD-UY: Detailed household electricity consumption dataset of Uruguay

Juan Chavat*, Sergio Nesmachnow*, Jorge Graneri*, and Gustavo Alvez**

*Universidad de la República, Uruguay, {juan.pablo.chavat,sergion,jgraneri}@fing.edu.uy
**UTE, Uruguay, galvez@ute.com.uy

ABSTRACT

This article introduces a data set containing electricity consumption records of residential households in Uruguay (mostly in Montevideo). The dataset is conceived to analyze customer behavior and detect patterns of energy consumption that can help to improve the service. The data set is conformed by three subsets that cover total household consumption, electric water heater consumption, and by-appliance electricity consumption, with sample intervals from one to fifteen minutes. The datetime ranges of the recorded consumptions vary depending on the subset, from some weeks long to some years long. The data was collected by the Uruguayan electricity company (UTE) and studied by Universidad de la República. The presented data set is a valuable input for researchers in the study of energy consumption patterns, energy disaggregation, the design of energy billing plans, among others relevant issues related to the intelligent utilization of energy in modern smart cities.

Background & Summary

Worldwide, electricity consumption of residential household showed an uninterrupted growth in the last fifty years[1]. It is expected that in 2050 the demanded electricity consumption doubles the one recorded at 2010[2]. Providing the future demanded electricity supply is a challenge and many investigations are taking place in this sense[3–6].

In Uruguay, electricity is provided by the state-owned electric company, Administración Nacional de Usinas y Trasmisiones Eléctricas (UTE). Uruguay has been recognized as one of the top countries with the most developed and used renewable energy sources. Uruguayan population is 3.4 million people, 1.3 million of them living in its capital, Montevideo. Electrification is considered universal, counting 99.8% of total areas (rural and urban)[7]. By July 2019, UTE provided electricity to 1,498,164 customers countrywide, of which 90.5% are residential[8]. The company provides a monthly average of 228 kWh to residential customers, 246 kWh in Montevideo and 216 kWh in the rest of the country[9]. In Uruguay, 87.3% of residential households have electric water heater (mainly for showers)[10]. The consumption of this equipment, which is fully manageable and has a high potential for thermal storage, represents approximately a third of the electrical consumption of all homes. In turn, the electrical matrix has diversified using renewable resources such as wind, solar, biomass energy, whose energy generation depends on weather conditions. This scenario allows implementing a proper management of the generated energy, making use of the potential of thermal storage in electric water heater of residential customers. A joint research project between UTE and the national university, Universidad de la República, was proposed to study the electricity consumption patterns of residential customers. In this context, the main motivations to create the presented dataset are related to study those patterns, detect similarities and anomalies, and be used as input of intelligent algorithms for planning, designing a recommendation system for citizens, and improve the overall quality of the electric service.

The systems designed to collect the data use different devices. The total household consumption is obtained from clamp meters or directly from smart meters (if available), while the dissaggregated consumption of the appliances is obtained by clamp meters or plug-in meters. Figure 1 shows a schematic overview of the data collection systems and the main processes involved.

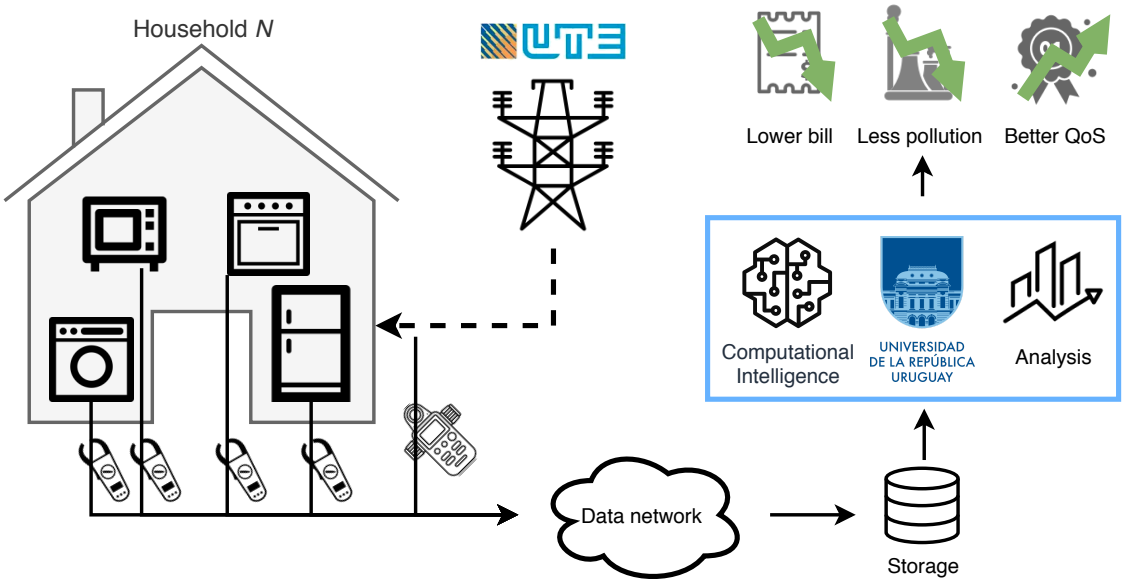


Figure 1. Schematic overview of the designed system for collecting and processing household electric consumption data.

The presented dataset, named *ECD-UY* after *Electricity Consumption Data set of UruguaY*, is divided into three subsets. The first subset consist of the total household consumption obtained from smart meters of 110952 customers countrywide, with a sample interval of fifteen minutes, for a period of 23 months. The second subset consists of the *electric water heater consumption* (disaggregated) of 268 households, from different cities in Uruguay, where 166 have customer information (i.e., the result of intersecting the households identifier of this subset with the customers information). The number of households that counts with aggregated and disaggregated records in this subset (i.e., the result of intersecting the households identifier of this subset with the total household consumption subset) is 135. The sample interval of the records is fifteen minutes for the aggregated consumption and one minute for the appliance consumption. The date range for the appliance consumption extends from 2^{nd} July 2019 to 26^{th} October 2020. The third subset consists of the aggregated records of nine households in Montevideo, and the disaggregate consumption of a set of appliances in each household (e.g., lamps, fridges, air conditioner, etc). The sample interval is one minute and the date range is from 27^{th} August 2019 to 16^{th} September 2019. Table 1 summaries the characteristics of each subset in the ECD-UY dataset.

subset	*households*	*aggregated*	*dissagregated*	*period*	*start date*	*end date*
total household consumption	110952	yes	no	15 min.	01/01/2019	03/11/2020*
electric water heater	268	yes**	yes	1 min.	2/07/2019	26/10/2020*
appliances consumption	9	yes	yes	1 min.	27/08/2019	16/09/2019*

Table 1. Summary of the three subsets contained in ECD-UY. * Periods may vary depending on the customer. ** Refers to the aggregated consumption present in the total household consumption subset.

Several datasets of energy consumption have been recently made available to the research community. Some well-known energy datasets are UK-DALE[11], including disaggregated consumption data from five UK households, REDD[12], including disaggregated consumption data from six households in New Jersey, USA, and AMPds2[13], including electricity, water, and natural gas consumption of a single house located in Vancouver, Canada. ECD-UY is the only public dataset describing residential electricity consumption with low-interval records in Uruguay, also the first available in its type in Latin America.

Methods

This section describes the methods applied for data collection, data communication, and pre-processing/cleansing. The information is reported for each collected subset.

Data collection

Different data collection processes were applied for each subset. The main details of the collection process, devices, and methods are reported next.

Total household consumption. Data of total household consumption collection was collected by the telemetry system of UTE. This system consists of smart meters installed in customers scattered around different Uruguayan cities, covering 31% of the total customers, at the moment of writing this article. The goal of the company is reaching a coverage of 100% of customers withing the next three years. The deployment of smart meters started in southern cities (Montevideo, Canelones, Maldonado, and Colonia), and continued to other cities along the country. As of June 2020, 245 000 smart meters have been installed. Actually, the installation of smart meters is part of the main operation of the company, within the development of a new smart grid infrastructure. Approximately 86% of the installed smart meters use the 3G network for transmitting the measured data. In turn, 10% use optic fiber communications and 4% use Power Line Communications (PLC).

The smart meters used in the deployment are KAIFA models MA110P (the most used, depicted in Figure 2), MA309P, and MA309D. All of them follow standards IEC 62052-11, 62053-11/21/23, and 62056-21/46/53/61/62. The devices allow measuring active and reactive energy, voltage and current, frequency, and offers 9600 bps modem communication, PLC/RF/GPRS/3G, RS-485. The measurement reporting period is configurable in ranges of 5, 10, 15, 30 or 60 minutes (the default value is 15 minutes). More specifications about these devices can be found at http://kaifametering.com/.

Figure 2. Smart meter Kaifa, model MA110P, installed by the Uruguayan electricity company, UTE (image by UTE, https://portal.ute.com.uy/medicion-inteligente).

The interval for data transmission was set to 15 minutes, the default value that KAIFA meters bring from origin, since this period is useful for characterization and billing purposes. This subset does not have the level of detail of the other two subsets (i.e., electric water heater consumption and disaggregated electricity consumption by appliance), which are obtained with a frequency of one minute, but it has the electricity consumption of at least ten times more dwellings. Gathering the total household consumption with smaller frequency (e.g., one minute) would imply handling with a very large volume of information. In turn, it would require a greater infrastructure of the company database. This has not been considered yet in the context of the pilot plan under development.

Electric water heater consumption. Electric water heater consumption data were collected from a set of 268 households of customers who were offered economic incentives (tariff reductions) to participate. The offered incentives were part of a commercial plan aiming to study electricity consumption patterns. The electric water heater appliance was chosen for the study because it is one of the most energy-intensive household appliances in Uruguayan households. The company invited customers to participate in a pilot plan, registering themselves to explicitly indicate their interest. The households of customers that accepted the participation were located in several country departments (Canelones, Montevideo, Salto, Paysandu, Maldonado, Rio Negro, Colonia, and San Jose) where the company installed a meter device and advised on the operation of it. The presented subset includes georeferenced consumption of customers located in three departments: Montevideo, Canelones and Paysandú.

Devices installed by UTE, used for measuring and transmitting the electricity consumption records, were smart switches (Sonoff brand, IM160810001 model) that consist of a plug-in power meter connected to the plug of the electric water heater. The used model measures a maximum wattage of 3500 W, voltage range between 90–250 V AC, a maximum current of 15 A, and its wireless frequency range is 80–160MHz under the standard IEEE 802.11 b/g/n.

Disaggregated electricity consumption by appliance. The data of electricity consumption by appliance was collected in a pilot plan developed in nine households located in Montevideo and Canelones. The monitored appliances vary between smart plugs, smart bulbs, motion and door opening sensors, thermostats, IP cameras, and gateways/hubs, this last used to communicate the measurements. The selection and labelling of the appliances were carried out by the occupants themselves. Thus, appliance selection criteria were totally under the control of the customers and not of the company. Three of the dwellings were apartments while the rest were houses. In average, the number of occupants per households was 3.0, 66.6% adults and 33.4% child. Clamp meters were used to measure the total aggregated consumption and the consumption of each monitored appliance, with a frequency of one minute. Table 2 summarizes the dwelling characteristics.

household id	*region*	*dwelling type*	*location*	*adults/childs*
1	Montevideo	House	Urban	2 / 0
2	Canelones	House	Coast	2 / 0
3	Montevideo	Apartment	Urban	2 / 1
4	Montevideo	Apartment	Urban	2 / 1
5	Montevideo	House	Urban	2 / 2
6	Canelones	House	Coast	2 / 2
7	Montevideo	House	Urban	2 / 2
8	Montevideo	House	Urban	2 / 1
9	Montevideo	Apartment	Urban	2 / 0

Table 2. Dwellings characteristics of the subset of disaggregated electricity consumption by appliances.

Communication

Collected data were transmitted to centralized data servers using different mechanisms. A description of the communication process for each subset is presented next.

Total household consumption. Once the aggregated consumption data is generated in each smart meter, it is transmitted to be stored in the Advanced Metering Infrastructure (AMI) of UTE. The AMI is a crucial component of modern smart grids, which is in charge of measuring the power consumption, implementing bidirectional communication between the customer and the service provider to communicate the obtained records, performing control tasks to optimize energy utilization, and implementing data management processes. The AMI is also the responsible of the communication with the smart meters and is the nexus with the billing system, the integrated operating system and the demand management system. Communication between the meters and the AMI is carried out via the 3G communication protocol for most of the households (93%). When the household location or dwelling makes it impossible to have a 3G connection (7%) (e.g., it is not in a coverage area), the smart meters are linked via RS-485 port to a hub connected to the fibre optic network of the National Telecommunications company (ANTEL). The architecture and processes of the communication system for the total household consumption subset is presented in Figure 3.

Electric water heater consumption. Subset records were collected using plug-in power meters and communicated by two different media: the household Internet connection or the 2G Internet connection. Communication media are described in the following paragraphs.

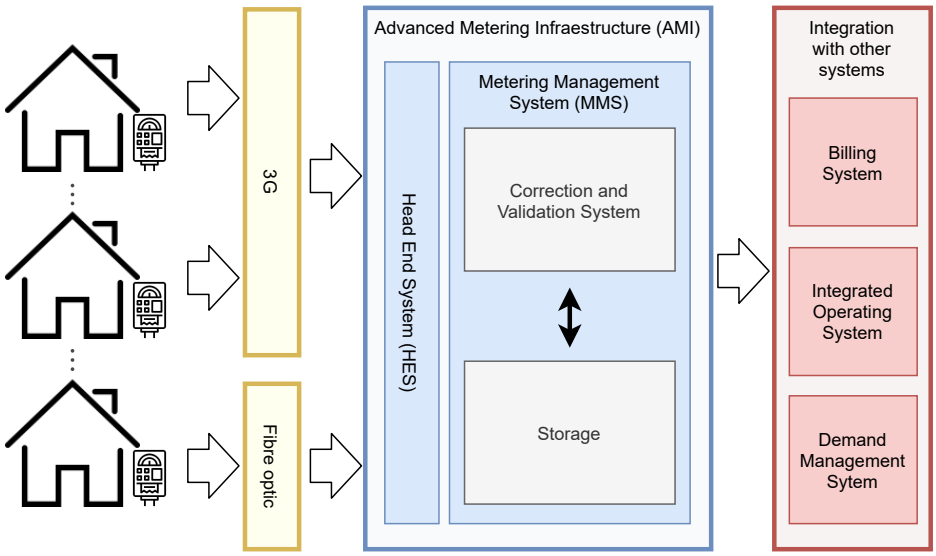


Figure 3. Architecture and processes of the communication system for the aggregate energy consumption data set.

Using the household Internet wireless connection, the plug-in power meters establish a bidirectional connection to send measurement data and receive switch-on/switch-off commands. Afterwards, the meters were updated to a model that allows establishing connections via a built-in 2G modem. The update of the meters improved the robustness of the connections, thus fewer data losing during the transmission, and brought independence from the Internet connection of the customers. Both connection media secure data by establishing VPN networks. The management of the communication was carried out by a chipset integrated into the meter itself. The model of that chipset is STM32 and the software embedded in it was implemented in C language and is property of UTE.

Collected data was transmitted with a frequency of one minute via the lightweight MQTT network protocol. Received data was processed at the UTE infrastructure using a demand management platform implemented over the Spring Boot framework and the Java programming language. The architecture and processes of the communication system for the electric water heater consumption subset is presented in Figure 4.

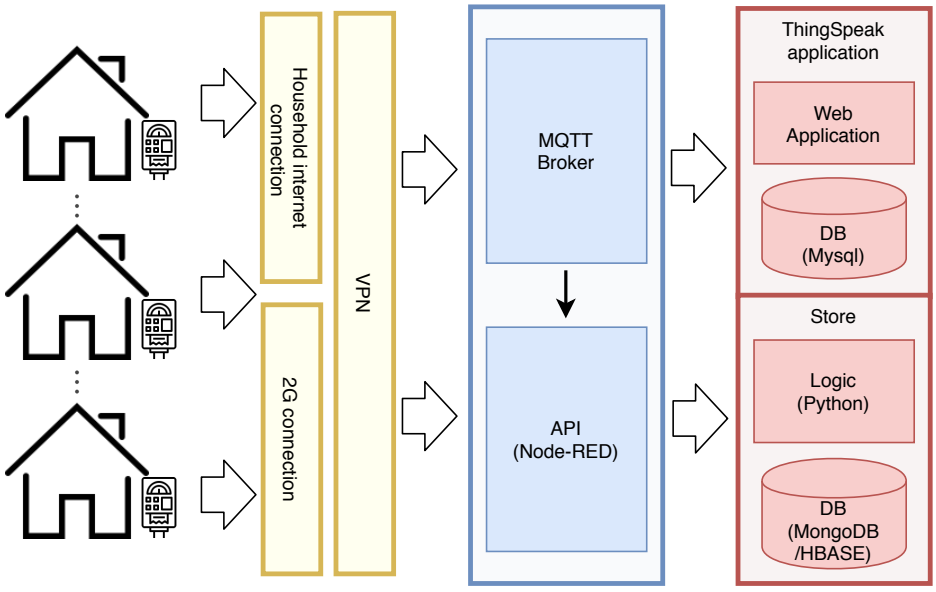


Figure 4. Architecture and processes of the communication system for the electric water heater consumption data set.

Disaggregated energy consumption by appliance. Clamps used for measuring electricity consumption communicate the collected recorded data to a gateway/hub inside the dwelling via the Zigbee 3.0 protocol. Once a measure is recorded, it is

sent via a wireless Internet connection to a remote third-party service. The service centralizes the storage of all data received from the clamps installed in the customers and it also associate the recorded consumption with dwelling metadata. The clamps measure and send the consumption with a sample period of one minute, but in case of loss of connection, the gateway/hub counts with buffer storage able to hold up to until the connection is reestablished. Regularly, UTE downloads the updated data, as text files, from the third-party servers.

Pre-processing and cleansing

Data were pre-processed and cleansed to be used as a common baseline for comparison of results in researches using the ECD-UY data set. Unix scripts using different tools (e.g., awk, split, sort, uniq, etc) and three Jupyter notebooks using Python language version 3 and several utility libraries, including Pandas, Numpy, and Dask[14,15], were implemented for the process. Data processing was performed using big data and urban informatics techniques, on the high performance computing platform of National Supercomputing Center, Uruguay (Cluster-UY)[16].

In general, transformation methods were applied over the collected raw data in order to: i) standardize the units, date formats, column names and file names, ii) remove unnecessary columns, and iii) build unique columns from the fusion of two or more columns. The transformations applied to customers and energy consumption data are described in the following paragraphs.

Customers data. The data about customers, provided by UTE, consisted of three files, one for each subset. The files contained customers that did not match with electricity consumption information, i.e., there was customers information without consumption records. Useless customers information was removed from each file and then, the three files were merged and standardised into a unique file. In the case of customers from the *dissagregated energy consumption by appliance* subset, the identifiers of the customers were changed in the consumption/appliances files to avoid collisions with different customers presented in the *total household consumption* subset. Also, the leading and trailing spaces in tension and tariff columns were removed.

Total household consumption. At the AMI module, the data was corrected and validated and then delivered to the meter data management system. Corrections were applied in case of detecting anomalies or missing data. After being processed by the AMI, the data was available for analysis. After applying the pre-processing and cleansing process to the raw data provided by UTE, the consumption records were stored in one file per month, to get appropriate file sizes (between 780 MB and 7.1 GB). The raw data had a file to describe the meter-customer relationship. After checking that there was only one meter per customer in the consumption files, the meter identifier was replaced by its corresponding customer identifier, reducing the overall size of the subset and simplifying the file structure. Finally, records with a null value in the customer identifier were removed, and the datetime column was represented in epoch time format, which allowed achieving a significant overall size reduction (20% less size, comparing with the version of ISO-8601 datetime format).

Electric water heater consumption. As part of the pre-processing of the collected records, the mean power and voltage of the electric water heaters were calculated as the average of all the measures within a minute, multiplied by a constant provided by the datasheet of the chipset (model H8012, included in the used meters). Also, the instant power was obtained by the time difference between two consecutive measurements multiplied by the same constant. Records processing was made by the same chipset in charge of the communication (model *STM32*) and the embedded software was implemented in C language. Then, the identifier of the meter was replaced by the identifier of the corresponding customer, in those records counting with customer information. Additionally, datetime columns were formatted to epoch time format. Both changes shorten the record length, reducing the overall file sizes. Finally, consumption records that belonged to customers/meters with less than 1440 records (i.e., the number of records corresponding to one day) were removed from the subset. Pre-processing and data cleansing reduced the total file size of the subset from 6.6 GB to 2.4 GB.

Disaggregated energy consumption by appliance. UTE collected customer information (e.g., household census areas and department) and related it to the appliances by a meter identifier. In turn, meter identifiers are the link between appliance information and its consumption records. Several types of consumption signals (e.g., active and reactive energy, active and reactive power, etc.) were recorded for a meter at the same datetime, as different rows in the consumption collection. In order to simplify the processing of multiple rows referencing the same meter at the same time, the multiple rows were transformed into a single consumption row with multiple columns, one per type of consumption signal. For the same reason, the consumption records of the appliances and the total consumption were separated into different collections.

During the cleansing stage, appliance information and consumption records were removed due to the meter identifiers were not present in both collections (i.e., appliances information without consumption records, or consumption records without appliances information). In total, 34 appliances and 1,163,714 consumption records were removed.

Code availability

Three Jupyter notebooks were implemented to facilitate the handling of the data set (one notebook for each subset). The notebooks are publicly available to download from https://github.com/jpchavat/ecd-uy. For a correct execution of the notebooks, Python version 3 and the Pandas and Numpy libraries are required.

Data Records

ECD-UY is available to download from the public repository https://bit.ly/3f3IVmK. The download file contains a structure with three directories, one per subset. The directory *total-household-subset* contains all the files related to the total consumption subset, *electric-water-heater-subset* contains the files related to the electric water heater subset, and *disaggregated-by-appliance-subset* contains the files related to disaggregated energy consumption by appliance. The data files inside each directory are in the CSV common format[17] and their columns are described in the following subsections. To reduce the amount of needed storage in the repository, large size files were compressed and presented together with reduced (and not compressed) sample of the data.

Customers information

The information of the customers is presented in a unique file, `customers.csv`, for all the subsets. The information consists of an identifier, characteristics of the electricity service contracted, and geolocalisation data in four levels. The records are detailed in Table 3. Customers records are related with the rest of the files by the value of the column `customer_id`.

`customers.csv`

id	*type*	*description*
`customer_id`	number	Unique value to identify the household
`tension`	string	Type of contracted tension, composed by a level mark (`BT` for low or `MT` for medium) and the voltage (e.g., `230V`, `400V`, or `15KV`)
`tariff`	string	Type of contracted tariff
`power`	number	Contracted power, in W
`department`	number	Department where the household is located
`section`	number	Censal section where the household is located
`segment`	number	Censal segment where the household is located
`zone`	number	Censal zone where the household is located

Table 3. Description of the records corresponding to the information of customers, present in the file `customers.csv`.

Total household consumption

The subset of total household consumption includes files of consumption records, one per month, each one named as `consumption_data_AAAAMM.csv`. The text "`AAAAMM`" included in the filenames corresponds to the year (`AAAA`) and the month (`MM`) of the records contained in the file. Table 4 reports the details of the information provided by each record. The information of the customers and the consumption records relate to each other by the value of the customer_id and id columns, respectively.

`consumption_data_AAAAMM.csv`

id	*type*	*description*
`datetime`	string	Datetime of the record, in Epoch time format
`id`	number	Unique value to identify the customer
`value`	number	Value of active energy, in kWh

Table 4. Description of records in files of the total household consumption data set.

Electric water heater consumption

The electric water heater consumption subset includes two files. On the one hand, the file `consumption_data_customers.csv` stores the consumption records of electric water heaters for which the customer information is available (stored in the customer information file, `customers.csv`). On the other hand, the file `consumption_data_timers.csv` stores the consumption records of electric water heaters without customer information. The conceptual separation into two files allows processing consumption data depending on the availability of information of customer, without requiring data filtering. A description of the records on each file is presented in Table 5.

`consumption_data_customers.csv`		
id	*type*	*description*
`datetime`	string	Datetime of the record, in Epoch time format
`id`	number	Unique value to identify the customer
`power`	number	Instant power in watts (W)
`voltage`	number	Instant voltage in Volts (V)
`consumption_data_timers.csv`		
id	*type*	*description*
`datetime`	string	Datetime of the record, in Epoch time format
`id`	number	Unique value to identify the timer (meter)
`power`	number	Instant power in watts (W)
`voltage`	number	Instant voltage in Volts (V)

Table 5. Description of records in files of the electric water heater data set.

Only the records of the file `consumption_data_customers.csv` are linked with the information of customers by the value of the column `id` in the consumption file, and the value of the column `customer_id` in the customers information file.

Disaggregated energy consumption by appliance

The disaggregated energy consumption by appliance data subset is integrated by the total aggregated consumption records plus the disaggregated consumption of different household appliances. The electricity consumption was recorded for the following appliances: air conditioner, dehumidifier, electric air heater, electric oven, electric water heatering, fridge, microwave, tumble dryer, and washing machine. It is worth clarifying that, since the customer chose the appliances to monitor, may occur that some of the appliances were present in the household but were not monitored. Three files are included in the subset: `appliances.csv`, `appliance_consumption_data.csv`, and `total_consumption_data.csv`. A description of the records in each file is presented in Table 6. The relationship between records of different files is given by the columns `customer_id`, `appl_meter_id` and `meter_id`, when applicable.

`appliances.csv`		
id	*type*	*description*
`customer_id`	number	Unique value to identify the customer
`appl_meter_id`	number	Unique value to identify the appliance meter
`appl_desc`	string	Appliance description (in Spanish)
`appl_type`	string	Appliance type (based on the nilmtk categories[18])

Continued on next page

Table 6 – *Continued from previous page*

`total_consumption_data.csv`

id	*type*	*description*
`datetime`	string	Datetime of the record, in ISO-8601 format
`meter_id`	number	Unique value to identify the appliance meter
`aenergy`	number	Active energy, in Wh
`aenergy_ph{1,2,3}`	number	Active energy in phase 1, 2 and 3, in Wh
`renergy`	number	Reactive energy, in VArh
`renergy_ph{1,2,3}`	number	Reactive energy in phase 1, 2 and 3, in VArh
`apower`	number	Aactive power in W
`apower_ph{1,2,3}`	number	Active power in phases 1, 2 and 3, in W
`rpower_ph{1,2,3}`	number	Reactive power in phase 1, 2 and 3, in VArh
`current_ph{1,2,3}`	number	Value of the current in phase 1, 2 and 3, in A
`pfactor`	number	Power factor (energy efficiency)
`pfactor_ph{1,2,3}`	number	Value of the current in phase 1, 2 and 3
`voltage_ph{1,2,3}`	number	Value of the voltage in phase 1, 2 and 3, in V

`appliance_consumption_data.csv`

id	*type*	*description*
`datetime`	string	Datetime of the record, in ISO-8601 format
`meter_id`	number	Unique value to identify the appliance meter
`aenergy`	number	Active energy, in Wh
`apower`	number	Active power, in W
`apower_ph{1,2,3}`	number	Active power in phases 1, 2 and 3, in W

Table 6. Description of records in the disaggregated energy consumption by appliance data set.

Technical Validation

This section describes sample experiments performed to support the technical quality of the ECD-UY dataset.

Total household consumption

The total household consumption subset includes the total aggregated consumption of 110952 households distributed in the 19 departments of Uruguay. On average, each household was monitored for 539.2 days and each day counts with 95.2 records.

Regarding the number of customers, the period of days monitored, and the number of records per day, two experiments were performed. The days considered for the experiments were classified into two groups according to the following completeness criterion. The expected number of records per day is 96 (i.e., one record every 15 minutes). The completeness criterion states that a *complete day* has at least 95% of the expected number of records.

Results of the experiment on the number of days per number of records indicate that more than 97% of the days have between 91 and 96 records, i.e. the vast majority of days meet the completeness criterion. Table 7 summarize the obtained results, disaggregated by the defined intervals on the number of records.

interval of records	*number of days*	*share*
(91, 96]	58122666	97.16 %
(86, 91]	355269	0.59 %
(81, 86]	92690	0.15 %
(76, 81]	289154	0.48 %
(72, 76]	76332	0.13 %
(0, 72]	887232	1.48 %
total	59,823,343	100 %

Table 7. Number of days per interval of number of records.

Regarding the number of customers and the number of days monitored, validation experiments were performed considering intervals of 60 days and 1 year. Using all the available days (not filtered by the completeness criterion), results showed that the 60-days interval with more customers ranges from 480 to 540 days and that 98.8% of the customers count with several monitored days in the yearly interval from 365 to 730 days. When considering only those days that meet the completeness criterion, the 60-days interval with more customers remains the same, as well as the yearly interval but with a share of 97.5% of customers. On average, the number of days per customer drops from 539.2 to 525.2 when filtering by the criterion.

In experiments using days that meet the completeness criterion, the total number of customers decreased from 110952 to 96565, mainly explained by a group of customers without even a day that meets the criterion. Detailed results on the experiments using 60-days intervals are reported in Table 8, and Figure 5 shows a histogram that relates the number of days and the number of customers. The table and the histogram shows, side by side, the results of the experiments when using all days and only those days that meet the completeness criterion.

interval of days	*all days*		*complete days*	
	customers	*share*	*customers*	*share*
(660, 690]	18400	16.58 %	0	0.00 %
(600, 660]	14105	12.71 %	14820	15.35 %
(540, 600]	21010	18.94 %	19032	19.71 %
(480, 540]	25643	23.11 %	26148	27.08 %
(420, 480]	18630	16.79 %	20069	20.78 %
(360, 420]	11870	10.70 %	13768	14.26 %
(0, 360]	1294	1.17 %	2728	2.83 %
total	110952	100 %	96565	100 %

Table 8. Number of days by customer, for all days and complete days (at least 95% of energy consumption records).

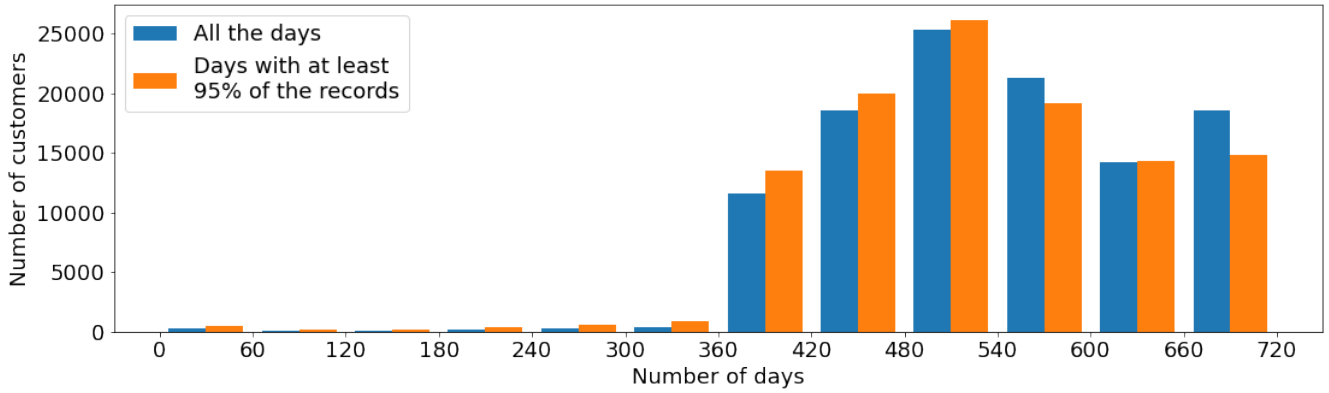


Figure 5. Histogram of the number of days with recorded consumption and number of customers for all days and complete days (at least 95% of records).

Figure 6 presents the mean energy consumption discretized in 15-minutes intervals, for four representative customers. Graphs show that the minimum consumption is during the night and some peaks are experienced mainly around the midday and at the end of the day. In turn, Figure 7 presents the mean energy consumption discretized in 15-minutes intervals for all the customers in the subset. The graph allows concluding that the minimum consumption occurs during late-night hours (around 4:00 AM). Two energy consumption peaks are recognised during the day, the lowest at midday, and the highest at around 9:00 PM.

Electric water heater consumption

The subset of electric water heaters consumption includes the disaggregated consumption records of electric water heaters. The technical validation evaluates the subset together with the corresponding total consumption, filtered by date range first and by the customer identifier then, resulting in the consumption of 135 customers.

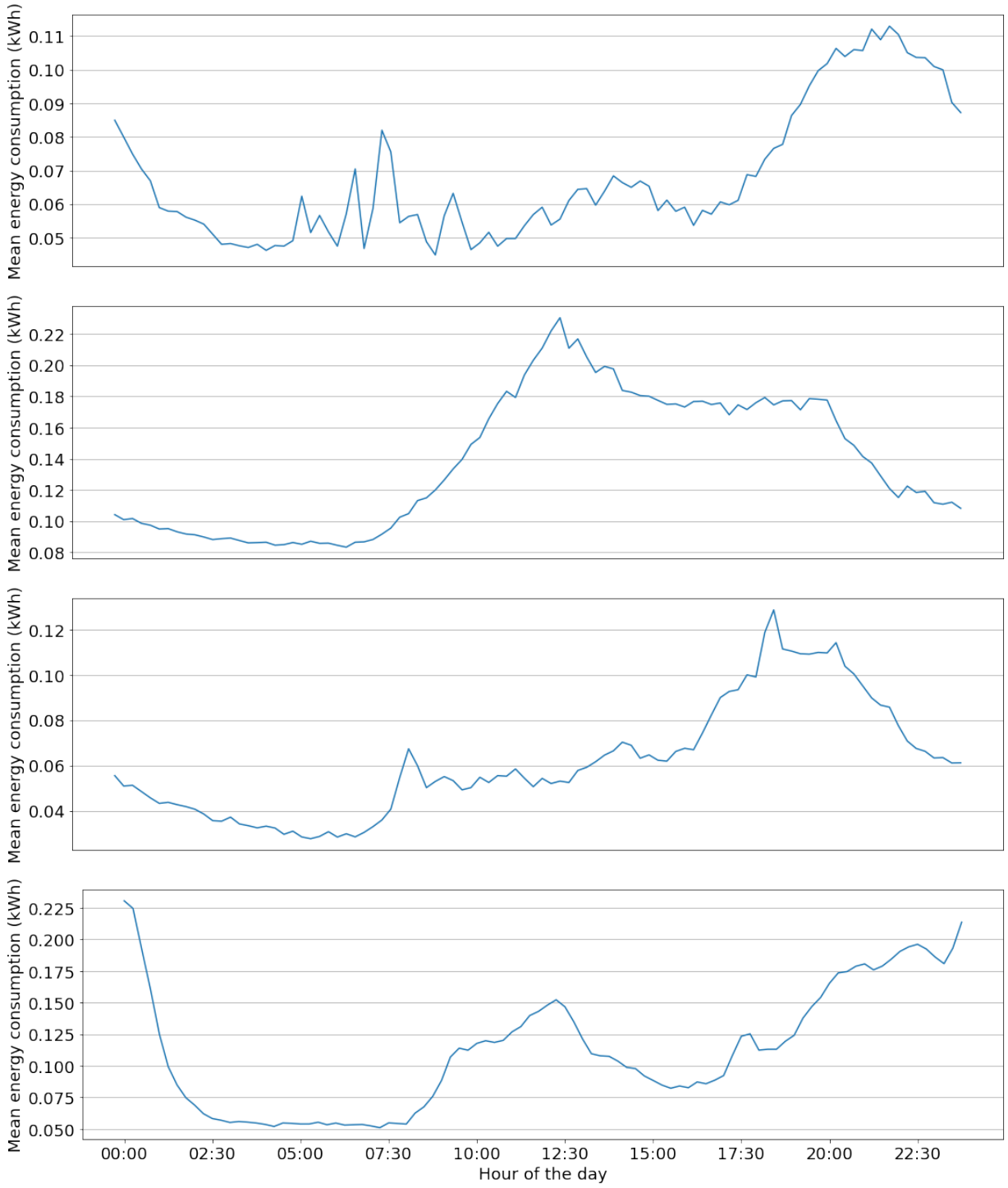


Figure 6. Mean energy consumption of one day for four customers (identifiers: #8037, #97875, #109846, #110088).

For the technical validation, records were filtered by percentile criteria trying to avoid data anomalies (e.g., exceptionally high consumption values). First, statistics and percentiles were calculated and studied to detect the outlier values, and then the consumption values detected as outliers were removed. In the case of the total consumption, no matter the customer, the records over a certain value were removed. For the disaggregated consumption, the percentile was calculated for each appliance and the records with a power consumption over the percentile were removed. Thus, the characteristics of each appliance (e.g., context, appliance model) were preserved.

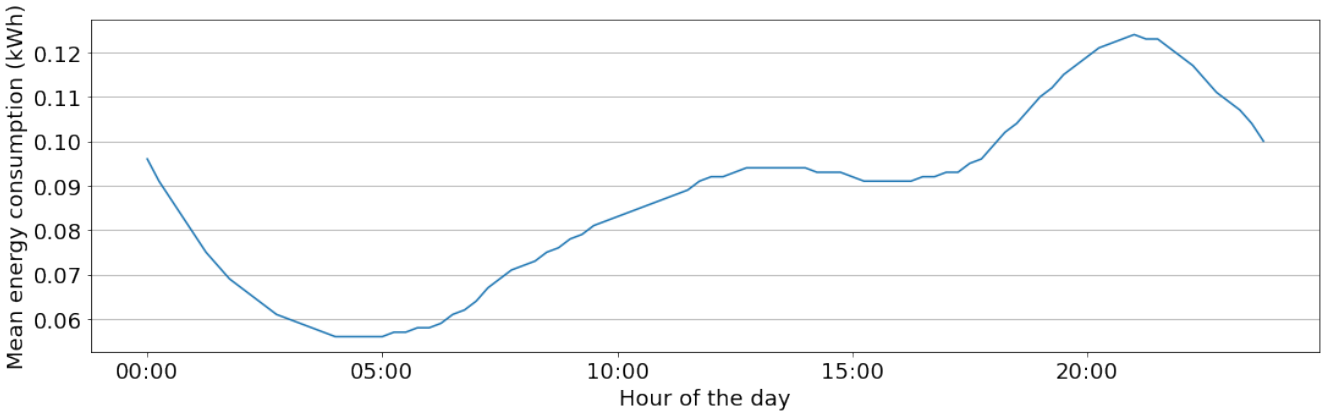
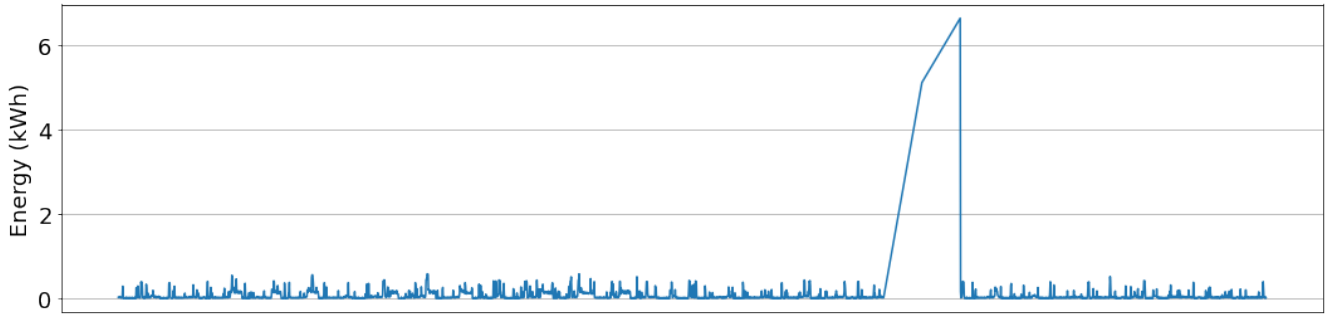
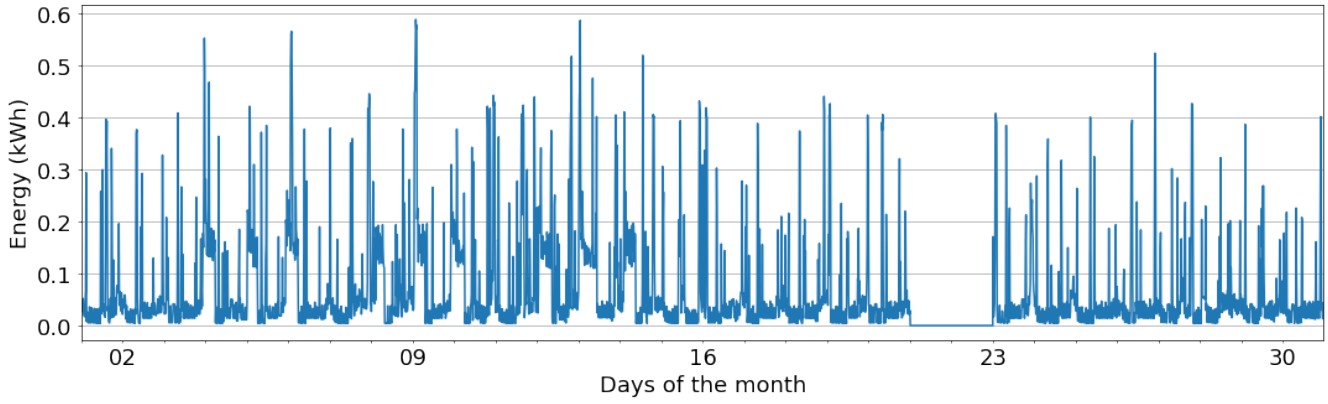


Figure 7. Mean energy consumption for a day.

Apart from removing outliers, the aggregated consumption was resampled to exactly 15 minutes and disaggregated data was resampled to one minute. Due to resampling, new records were created in periods of missing data. Both, records values over the criteria and records created after the resample, were refilled/filled with zero values. Figure 8 shows the aggregated consumption of one month (September 2019), for one customer (#69806), before and after filtering, resampling, and refilling the records. That is a case where data was missed for a long period. Instead, only two exceptionally high values were recorded. The missing period and the between values were refilled/filled with zeros.



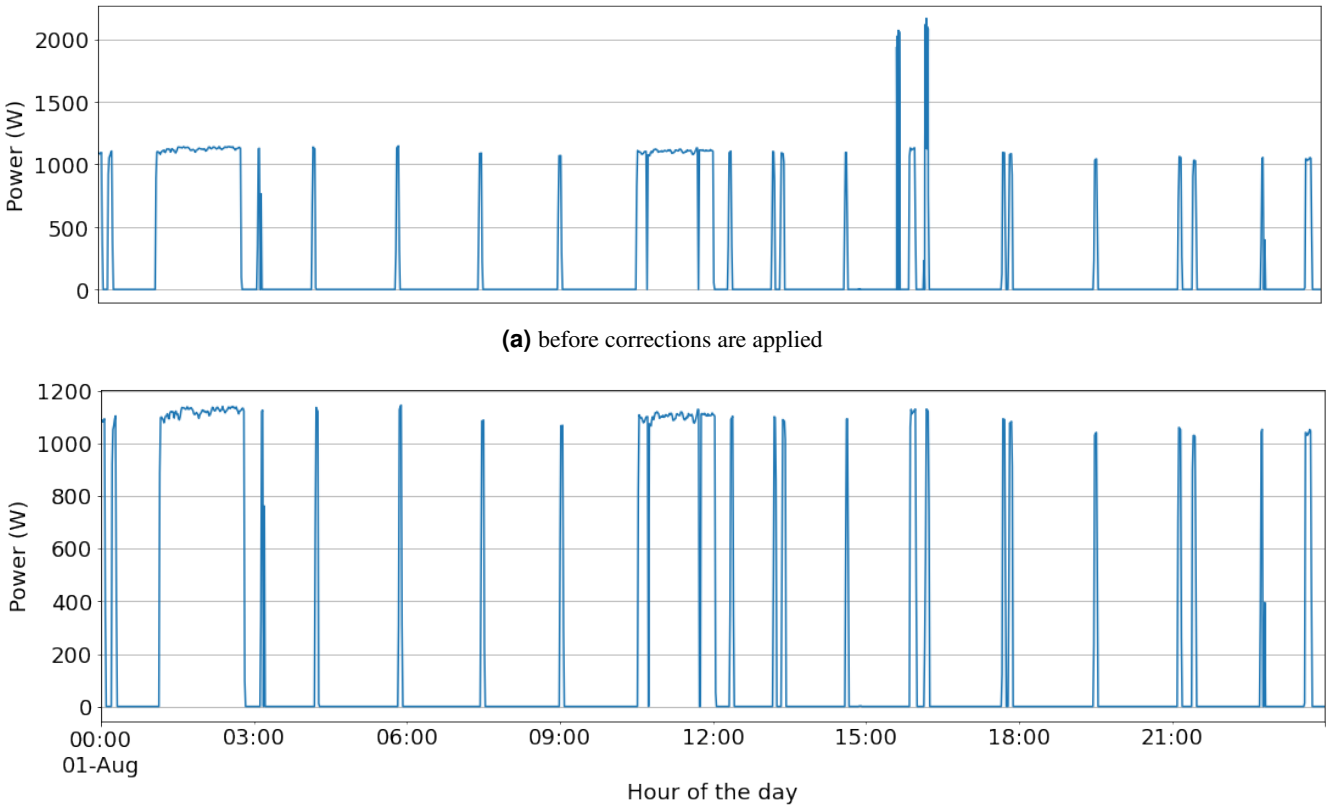
(a) before corrections are applied



(b) after corrections are applied

Figure 8. Example of one month of total household consumption with outliers and missing values, and its subsequent correction, for customer #69806.

Likewise, Figure 9 shows the disaggregated consumption of one representative day (August 1st, 2019), for one customer (#115609), before and after filtering, resampling, and refilling the records. In this case, no data was missing, but exceptionally high values were recorded. The detected high values were set to zero.



(a) before corrections are applied

(b) after corrections are applied

Figure 9. Example of 24 hours of electric water heater consumption with outlier values and its subsequent correction, for customer #115609.

Depending on the data purpose of use, anomalies can be ignored or treated in different ways. The decision on how to treat the data was left to the final user, so the result of the data cleaning described for this technical validation was not included as part of the presented data set.

After data cleansing, the mean power consumption was calculated for periods of 15-minutes in a day. Results showed that the electric water heater has the most relevant share of the total household consumption. On average, it represents 27% of total consumption, reaching 35% during peak hours. Figure 10 shows the mean total and electric water heater power consumption for a day, highlighting the important contribution of electric water heater consumption to the total consumption.

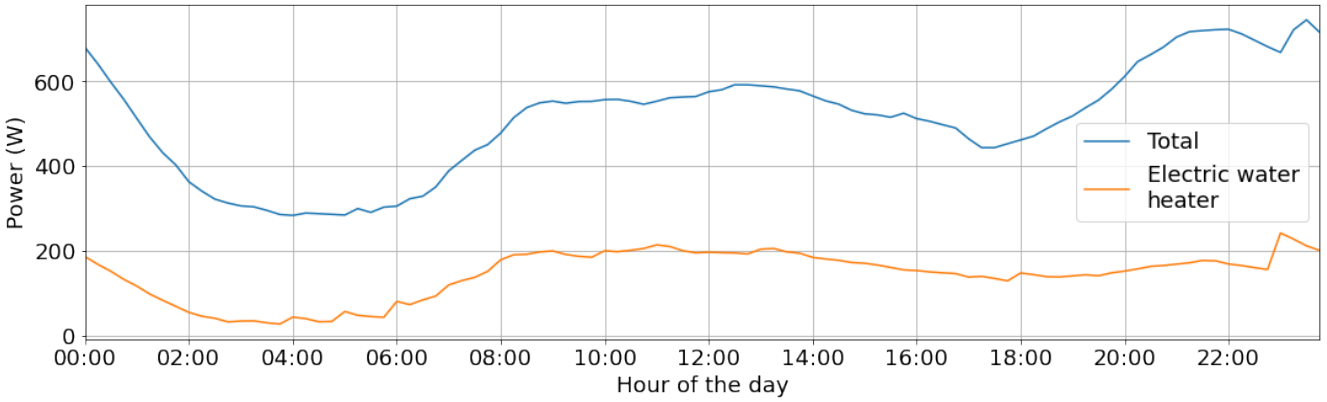
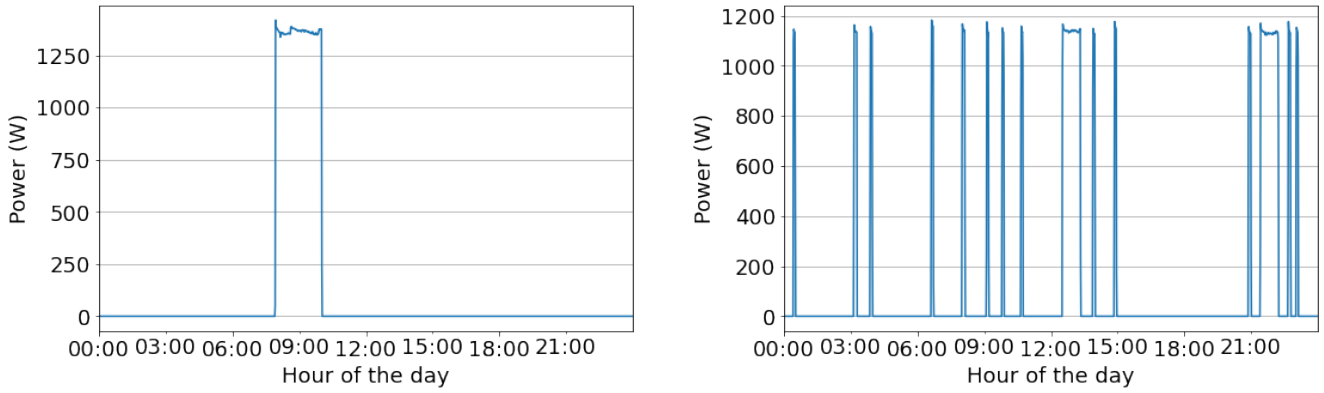


Figure 10. Mean power consumption for the total and electric water heater consumption of a day.

Regarding the consumption of the electric water heater, two basic patterns were identified. The first pattern shows appliances on only for those moments when hot water is needed (e.g., shower). This case is observed in the sample graphic in Figure 11a. This pattern is also related to households with highly efficient electric water heaters appliances, which avoid standby losses. The second pattern is related to electric water heaters in standby mode during all day, with periodic consumption peaks. The water heater is automatically switched on several times a day (for short periods of a few minutes) to preserve water temperature. Figure 11b presents a sample consumption graphic for a water heater that meets this consumption pattern.



(a) appliance switched on only for specific uses (customer #65374, November 2^{nd}, 2019)

(b) appliance in standby mode all day, switched on when needed (customer #4729, October 30^{th}, 2019)

Figure 11. Example of electric water heater consumption patterns.

Disaggregated energy consumption by appliance

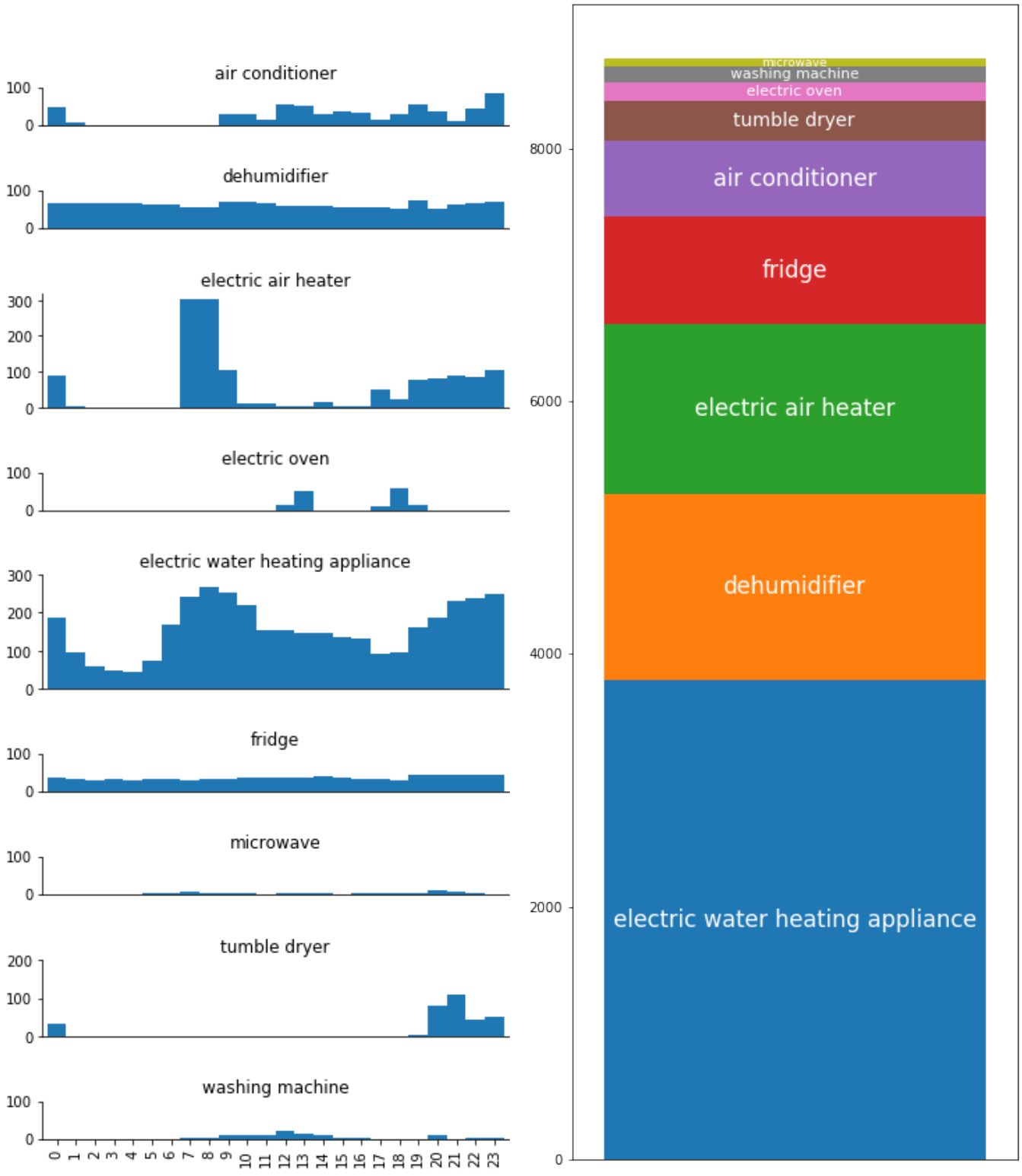
Recording periods of household appliances consumption lasted on average 19 days. During that time, data gaps (consecutive missing records) and outlier values were recorded, mainly due to meter failures and and connection issues. Gaps duration depended on the appliance, ranging from one to almost seven hours. Table 9 presents detailed information about the recording and gaps duration, disaggregated by appliance.

appliance	*# appliances*	*mean # gaps*	*mean gap duration*	*mean recorded duration*
electric air heater	2	177.0	01:50:59	19 days 19:05:00
electric oven	1	120.0	01:07:33	19 days 19:05:00
tumble dryer	1	116.0	01:40:46	19 days 19:05:02
washing machine	5	114.0	02:17:09	19 days 16:07:59
electric water heater	5	107.6	01:30:54	19 days 19:04:59
microwave	4	104.5	01:37:33	19 days 19:04:59
fridge	5	81.2	06:38:48	19 days 13:05:59
dehumidifier	1	74.0	03:25:36	18 days 04:41:58
air conditioner	3	73.3	01:50:51	19 days 19:04:59

Table 9. Detailed information about the recording duration and data gaps.

To avoid including gaps and outliers during the validation, data were filtered, resampled and refilled as described below. First, all the consumption values lower than zero or greater than the 99^{th} percentile were set to zero. Percentiles were calculated for each appliance, and not in general, to preserve the household context and the characteristics of each appliance. Then, consumptions were resampled to one-minute periods, resulting in many new records where data were missing. Finally, the value of the new records was set to zero.

Figure 12 shows information about appliances power consumption. Figure 12a shows 1-hour histograms of the mean power demanded by each appliance. For some appliances, specific operating times are identified, e.g., for the electric oven and the washing machine, whereas for other appliances, the demand is almost constant during the day, e.g., for the fridge and the dehumidifier. Figure 12b shows a stack bar that accumulates the mean power demanded by each appliance in a day. The appliances with the greatest impact on total demand are the electric water heater, the dehumidifier, and the electric air heater.



(a) Histograms of mean power demanded per hour by each appliance. **(b)** Stacked bar of mean electricity demanded by each appliance.

Figure 12. Mean power demanded by hour (histograms) and in a day (stack bar), for each appliance present in the subset.

Usage Notes

Any software that handles CSV files can load the ECD-UY data set. In the presented article, for processing the dataset the software used was Python version 3 and the libraries Pandas and Numpy. Loading big size files entirely in RAM memory may cause several problems that can be avoided by using the library Dask[15]. Dask can execute operations in parallel and load just the necessary data in memory. Depending on the type of processing, it may be useful to transform the dataset from CSV to Apache Parquet format[19], which is a structured, column-oriented, compressed and binary file format that can be used for efficient processing data in Apache Hadoop and similar frameworks.

For the previously described data gaps, a rule of thumb is suggested to classify between short and long ones. If the gap duration is lower than 12 minutes, it may be considered short, elsewhere, it may be considered long. Short gaps are likely to be refilled by a method, e.g., interpolating or averaging the previous and forward value records. Long gaps can be considered as long periods of the appliance switched off, and therefore it would be correct to assign zero value to its consumption.

For analyzing irregular recording periods, a resample process may be applied together with value refilling criteria. A suggested resample/refilling criterion consists of creating records with regular periods and refilling with the maximum or average value (values) present in each regularized interval. An example of a resample/refilling process is implemented in the Jupyter Notebook corresponding to the technical validation of the Electric water heater subset, available to download at `https://bit.ly/3yv5Hvc`.

References

1. International Energy Agency. World Energy Outlook 2020. White paper, https://www.iea.org/reports/world-energy-outlook-2020 (May 2021) (2020).
2. Larcher, D. & Tarascon, J. Towards greener and more sustainable batteries for electrical energy storage. *Nat. Chem.* **7**, 19–29 (2015).
3. Ford, R. *Reducing domestic energy consumption through behaviour modification*. Ph.D. thesis, Oxford University (2009).
4. Luján, E. *et al.* Cloud Computing for Smart Energy Management (CC-SEM Project). In *Smart Cities*, vol. 978 of *Communications in Computer and Information Science* (Springer, 2019).
5. Orsi, E. & Nesmachnow, S. Smart home energy planning using IoT and the cloud. In *IEEE URUCON* (2017).
6. Chavat, J. P., Nesmachnow, S. & Graneri, J. Non-intrusive energy disaggregation by detecting similarities in consumption patterns. *Revista Fac. de Ingeniería Universidad de Antioquia* (2021).
7. Tasa de electrificación urbana y rural. Tech. Rep., MIEM (2018). https://www.miem.gub.uy/energia/series-estadisticas-de-energia-electrica (in Spanish, May 2021).
8. Número de clientes de energía eléctrica por sector. Tech. Rep., MIEM (2019). https://www.miem.gub.uy/energia/series-estadisticas-de-energia-electrica (in Spanish, May 2021).
9. Memoria anual 2017. Tech. Rep., UTE (2019). https://portal.ute.com.uy/sites/default/files/generico/Memoria_2017.pdf (in Spanish, May 2021).
10. Microdatos de la encuesta continua de hogares. Tech. Rep., Instituo Nacional de Estadística, Uruguay (2019). https://www.ine.gub.uy/encuesta-continua-de-hogares1 (in Spanish, May 2021).
11. Kelly, J. & Knottenbelt, W. The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes. *Sci. data* **2**, 1–14 (2015).
12. Kolter, J. & Johnson, M. J. Redd: A public data set for energy disaggregation research. In *Workshop on data mining applications in sustainability*, vol. 25, 59–62 (2011).
13. Makonin, S., Ellert, B., Bajić, I. & Popowich, F. Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. *Sci. data* **3**, 160037 (2016).
14. McKinney, W. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython* (O'Reilly Media, Inc., 2012).
15. Rocklin, M. Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th Python in Science Conference*, 130–136 (2015).
16. Nesmachnow, S. & Iturriaga, S. Cluster-UY: Collaborative Scientific High Performance Computing in Uruguay. In *Communications in Computer and Information Science*, 188–202 (Springer, 2019).
17. Shafranovich, Y. RFC 4180: Common format and mime type for comma-separated values (csv) files. *The Int. Soc.* **54**, 258 (2005).

18. Batra, N. *et al.* NILMTK : An Open Source Toolkit for Non-intrusive Load Monitoring Categories and Subject Descriptors. *Int. Conf. on Futur. Energy Syst.* 1–4 (2014).

19. Vohra, D. Apache parquet. In *Practical Hadoop Ecosystem*, 325–335 (Springer, 2016).

Acknowledgements

This research was partly supported by CSIC-UDELAR and UTE through project “Computational intelligence for characterization of electric energy consumption in residential households". The work of S. Nesmachnow is partly supported by ANII and PEDECIBA, Uruguay. The publication fee is partly supported by PEDECIBA and CSIC-UDELAR, Uruguay, and personal funds.

Author contributions

Juan Chavat: Data analysis, data cleansing, manuscript writing. Sergio Nesmachnow: Data analysis, manuscript writing. Jorge Graneri: Data analysis. Gustavo Álvez: Data analysis, manuscript writing.

Competing interests

Authors declare no conflict of interest.