Energy Informatics

DATA NOTE

Open Access

Energy consumption data collection: case study on data center in a Thai University

Withit Chatlatanagulchai[1†] and Chantana Chantrapornchai[2*†]

†Withit Chatlatanagulchai and Chantana Chantrapornchai contributed equally to this work.

*Correspondence: fengcnc@ku.ac.th

[1] Department of Mechanical Engineering, Kasetsart University, Ngamwongwan Rd., Bangkok 10900, Thailand
[2] Department of Computer Engineering, Kasetsart University, Ngamwongwan Rd., Bangkok 10900, Thailand

Abstract

Objective: Energy usage in has been increased due to the rising demand of cloud infrastructure. The government policy has been focused on building the green IT data center. The energy data need to be collected in order to monitor the energy usage. However, in an old typical data center, the building has been built with no support of such data collection. In this research, we aim to design the energy data collection system for our existing data center, a case study of data center in Thailand at the university. Based on the collected data, an energy usage monitoring system and prediction can be developed.

Methods: In the case study of Kasetsart University data center, the building and electric layouts were predetermined. The building layout and existing IT hardware were investigated. We designed the meter types and the number of meters to be installed for the building the energy data collection system. The corresponding database system was also designed for data logging, data visualization and analysis purpose.

Results: As a result, 25 installed meters along with the add-on network system were installed for logging data. A data usage example was demonstrated by building the data visualization and analysis. The presented 1 year dataset collected showed the changes of energy usages which can be used to compare with real activities happening in the campus. This encourages the integration of other related environment data such as outside temperature which may affect the electric billing cost. The dataset can be used for prediction of the electric usage; thus, the policy for reducing the electric billing cost could be established. In this paper, as a data note, we focus on the methodology of data collection required for data center.

Keywords: Energy consumption, Data center, Energy dataset

Introduction

Data center is one of the top organizations that consume energy. From Sverdlik (2016), data centers in the United States consume upto 70 billion kWh which is about 2% of the country's energy usage and equivalent to total 6.4 million household energy usage. This consumption rate has increased from 2010 to about 4%.

Recent previous works reported that most energy consumed are from servers and cooling systems (Hemsoth 2016; Jennings 2016; Office of Energy Efficiency and Renewable Energy 2016). The expense is directly related to the server about 60% and the cooling system about 40%. Also, from UN report, the cooling and heating systems are sources of energy usage upto 70% of the city (Zorba 2024).

In a typical data center, there are various sources of power consumption including IT and non-IT data components. Efficient power usage in a data center tends to spend most of the power in IT equipment. However, this is hardly possible in the tropical country like Thailand. Several servers running computing-intensive services tend to consume lots of energy and yield lots of heat dissipation. Proper cooling equipment known as CRAC (Computer Room Air Condition) are usually required to cool down the servers so that they can run services normally. However, the size of CRAC system and the operational policy in a data center is varied depending on many design factors, e.g, vendor setup, floor plan layout etc.

In this research, we aim to develop an energy usage monitoring system and prediction, a case study of data center in Thailand at the university. The building was constructed several years ago. The physical layout as well as the rack locations were fixed. The servers needed to run 24/7 and there were no programs for resource usage profiling. To achieve our goal, we study existing layouts and IT facilities and attempt to design the appropriate meter installation solutions under the current situation for collecting energy usage data under the reasonable budget. The network wiring and component connection were designed for data logging. At last, the obtained data were analyzed and the prediction system were developed. The energy usages were collected as an open data collection (Chantrapornchai and Chatlatanakulchai 2023).

This presented data note demonstrates the contributions in the following aspects.

- Experience sharing in designing the system for energy data collection in the fixed floor plan data center.
- Dataset and example visualization aided in finding the factors that affect the power usage in the data center.
- Demonstration of an application of machine learning to create the prediction models.

Several datasets have been proposed about energy consumption. For example, in data.gov (as of 28 January 2024), we found 9 datasets related to energy consumption. They included resident energy consumption, monthly and annual energy consumption, building synthetic dataset, etc. Some of them were the table reports and presented APIs for data visualization. From data.world, there are 81 related datasets by searched keywords “data center energy consumption” (as of 22 March 2024). Only a few of them related to electricity power consumptions. In the following, we give some examples of detailed datasets that are available for data scientists for in-depth analysis.

1. Dua and Taniskidou (2017) presented the energy consumption in households divided into room types and three submeters. It consists of watt-hr energy consumption values. The total period is 47 months. The collected data contain 9 fields including datetime, voltage, and global reactive power. The dataset was also published in Kaggle (https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set).

2. Candanedo et al. (2017) collected energy usage in a household. The sampling time is 10 seconds and the total period is 4.5 months. The temperature and humidity data were also collected every 3.3 min. The dataset includes the weather data collected from the base station at Chievres Airport, Belgium.
3. Makonin et al. (2016) presented the dataset of households during 2012-2014 in Canada. The meters were installed for each room in the house. The hourly weather data were also collected. The dataset contains CSV file of electric meter data, weather data, historic electric, water, gas billing and usage.
4. Chavat et al. (2022) presented the energy consumption of households in Uruguay. The dataset was collected by Uruguayan Electricity Company (UTE) and was studied by Universidad de la República. There are four subsets: total household consumption, total electric water heater, electric appliance, and customer information.
5. Bazurto et al. (nd) presented the dataset from the facilities of the Information Technology Center (CTI) of the Escuela Superior Politécnica del Litoral (ESPOL). The data were collected from an HP Z440 workstation for 245 days (35 weeks). The sampling interval was one value per second. The dataset contains attributes for the power consumption of CPU, GPU, along with the memory usage, as well as the CPU temperature etc.
6. Sheppy et.al. Michael et al. (2011) collected the dataset for NREL building. NREL building was a research facility building. The dataset contains hourly data for: Total Cooling (kW)-Total Heating (kW)- Total Mechanical (kW)-Total Lighting (kW)-Total Plug Loads (kW)-Total Data Center (kW)-Total Building (kW)-PV (kW)-Building Net (kW).

After achieving the dataset, machine learning or AI is often utilized for predicting the energy consumption usage (Gao 2018; Li et al. 2017; Shoukourian et al. 2017). Some famous current data center has utilized AI in some form for managing facilities (Donovan 2018; DeepMind 2018).

From previous work, two aspects of the energy datasets were demonstrated. First, in the aspect of using meters, datasets were proposed in the context of household building since the meter installation logistics was not too complex. Second, in the context of server loads, CPU/GPU workload along with the power consumption on the computer server were gathered. Most datasets report only based on a specific computer since since this approach needs some middleware programs to do profiling which cannot be done easily in a real data center due to several reasons. For example, the data logging may degrade server performance and the server could not be stopped to install the program, etc.

This data note mainly differs from previous datasets in the aspect of the organization context. We rely on the data center context, rather than the household. The data center is divided into rooms containing servers and IT equipment. In achieving the data, challenges were relevant to data logger installation in the building that there had a fixed layout and the operations that run continuously. Due to the management policy, we cannot stop the operations and installed the profiling software. We, then, had to focus on the design of data logger installation which needs to solve the solutions such as some meters exist; however, data cannot be logged. Or some portion of the floor plan where the servers are located did not have separate meters while considering the budget. As a results, it enables us to collect power usage for servers as well as estimate the power usage for other IT components in the data center. The design experience along with the datasets will be benefits to the readers in the fields.

In the next section, methodology of acquiring data is presented. Section Results presents the data results and section Data processing and analysis demonstrates the usage of data. Section Discussion and implications discusses the results and implications.

Methodology

Organization context

Our university, Kasetsart University, is a large-size university, serving around 38,000 students. The data center is under Office of Computing Service (OCS) which serves the core functions of the university such as the registration, payroll, e-mail, cloud service, etc. The data center occupies one floor of the OCS building. The building also includes offices, lecture rooms, computer laboratory rooms, conference rooms, etc. According to the university billing, the monthly billing cost for the data center is around 40% of the whole building.

From the previous record, e.g. January, 2017, the total electricity cost of the building was 500,000 THB Off-Peak Power was 370 kW and On-Peak Power was 357 kW which is about 37% is the total cost. The expense is higher around 10–15% during March, April, May for the summer time in Thailand.

For a typical university in Thailand, the data center is medium size. It is divided into several rooms. Each room contains servers of the university and colocated servers. The main servers of the university have the workload of registration, personnel, payroll, etc. Our university was issued the funding to the university to develop the energy monitoring use case in a data center of the university under the theme of big data analytics for energy, by Energy Policy and Planning Office (EPPO), Ministry of Energy, Thailand, in 2017. To the best of our knowledge, we were the first university granted by EPPO to develop the power consumption monitoring for data center context.

In order to proceed, power usage must be logged. Proper meters and data loggers must be properly installed. Challenges in our context are the following:

- The layout of the server and CRAC locations which cannot be changed and the servers cannot stop running.
- There are many IT and non-IT equipment which may directly or indirectly be connected to the meters.
- Fixed amount of budget e.g. 500,000 THB was for the hardware cost.
- Some equipment are modern; i.e. there are power usage display on them.

The following subsections describe our process to tackle these challenges. We have to analyze the floor plan, electric wiring, and study of existing meters. Then, within the budget and criteria, meter installation plan is designed. At last, the data integration is presented.

Floor plan analysis

The whole building serves many functions including office administration, computer training rooms, lecturer rooms, etc. It has the power meter installed by the Metropolitan Electricity Authority on the 1^{st} floor. Figure 1 shows the floor plan of the data center on the 6^{th} floor. The MDB room is the controlled electricity room connected to the 1^{st} floor main electricity room. There are 5 rooms divided to store servers, networks, equipment: server room 1, server room 2, co-location room (servers not own by the university), NOC room and network room. Each room is connected with ULC channels for power meters.

The server room1 and server room 2 shown in Fig. 1 are used for the servers of university. For example, server room 1 shown in Fig. 1 is where the main servers of the university are located. This room is connected to CRAC3 and CRAC4. CRAC3 and CRAC4 were alternately used daily. The temperature setting is 21 centigrade. The ULC5 and ULC6 control boxes are connected for both server racks for the server room 1. ULC9 and ULC10 are in the colocation rooms. ULC1 and ULC2 are in the spare room which connects to the reserved racks. ULC3 and ULC4 are in the network room. All CRAC1-CRAC6 are central cooling systems where each two are connected to blow the cool air through the underground of the aligned server rack in the room and the hot air is blown out from each room in the pipes hidden in the ceiling.

As shown in the previous work, reducing the energy usage can be related to the air flow management and other IT equipment. The effective cooling and airflow in a data center is necessary to eliminate the heat (Phelps 2018). The main concept is that cool air from CRAC should go through the server intakes and hot air from the server release goes through CRAC returns. The CRAC setup should be perpendicular to the rack rows. In our case, due to the fixed infrastructure in our data center, it is not possible to move the server racks. However, our CRAC was already setup according to this manner. We have surveyed the number of servers used and IT equipment in the data center room. The main energy consumption in the room were from the servers and CRACs. The running servers already connected to ULC channels which can later be connected to the meters while other equipment like network switches, UPS, storage, etc. cannot be

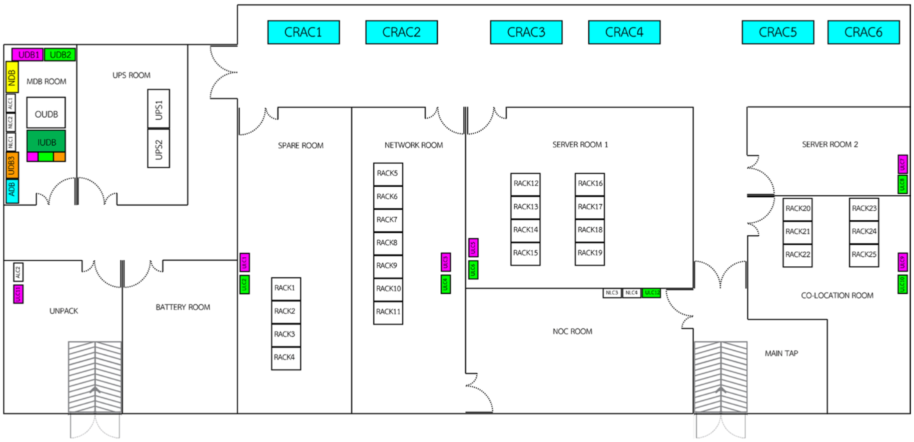


Fig. 1 The whole data center room layout

monitored directly. Since the cost of hardware installation is high if more meters and sensors were used. The power consumption of these IT equipment may be inferred from other meters.

From the layout, the sources of energy inefficiency may be from the following. (1) the CRAC temperature setting was fixed. There are temperature sensors in the ceiling of the room but the temperature values in the room do not use to control the CRAC setting (Phelps 2018). (2) there were some equipment that was not in use but not turn off, such as UPS, network storage, etc. (3) the CRAC temperature setting can be varied depending on the server racks (Fukumoto et al. 2010). (4) the server racks were out-dated; the modern server racks that can monitor server load, resource management, may be utilized (Strom 2016). etc.

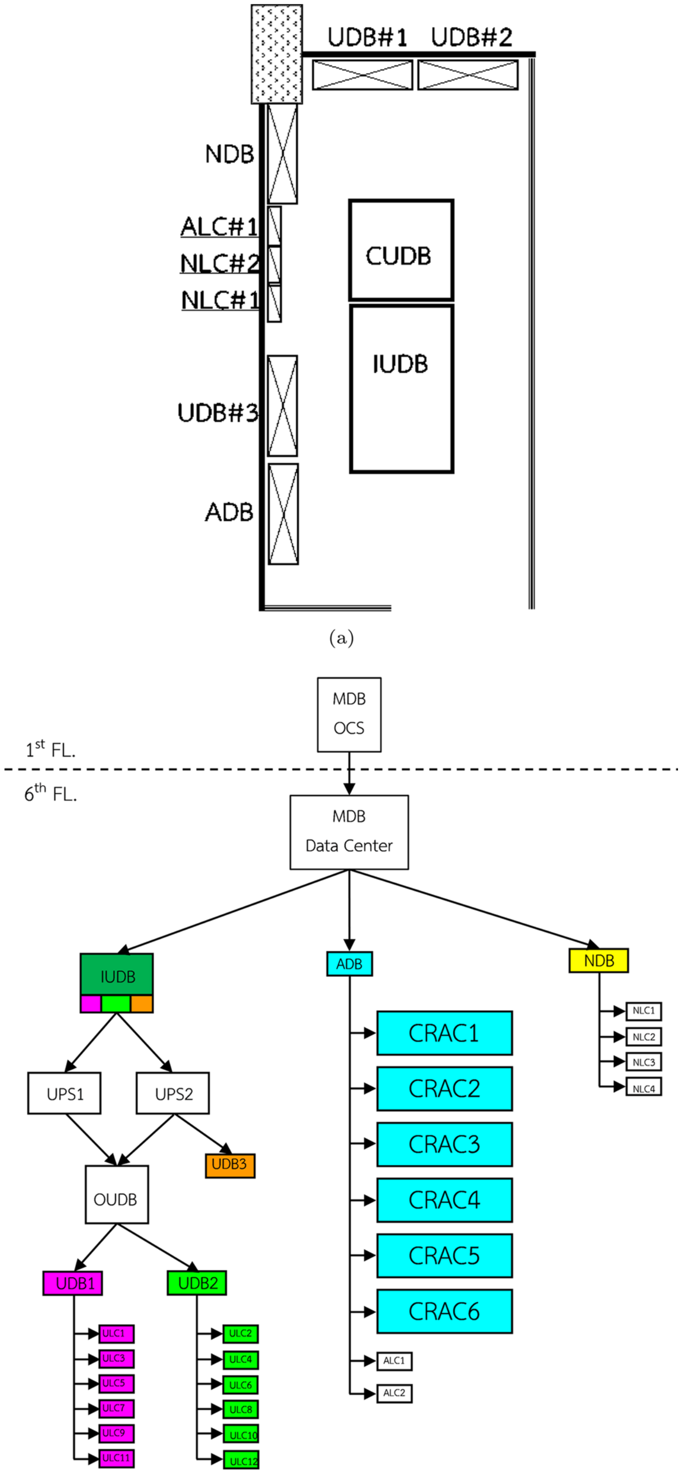
Main power consumption sources are from mostly servers, and CRACs, in the data center. We need to log the power consumptions for all these devices are much as possible. Since CRACs blow the heat from servers, induced by server workloads, CRACs' setting can be readjusted dynamically, making it suitable for the servers' heat dissipation. We need to measure the power usage of server loads which can imply the server workload. In order to achieve this, let us consider the electricity room shown Fig. 2a (called MDB room). The MDB room is the controlled electricity room. In this room, there are three main electric controls: IUDB, ADB, NDB. The electric connections are wired through UDB1 and UDB2 channels which is under OUDB and IUDB as in Fig. 2b. IUDB connects the two main UPS's (UPS1,UPS2) of the university servers which are both connected to OUDB. Then, it is split to UDB1 and UDB2 which are electric controls for server racks. UDB1 controls ULC1, ULC3, ULC5, ULC7, ULC9 and ULC11 while UDB2 controls ULC2, ULC4, ULC6, ULC8, ULC10, and ULC12 respectively. On the other hands, ADB connects directly to all CRACs. NDB connects to NLC1 to NLC4 for other devices such as networking. The IUDB, OUDB, UDB1, UDB2, ADB, NDB contained meters previously installed from the outsource vendors which can view the power usages for three-phase electric but cannot log data.

Meter installation

In the data center, there are several types of equipment, IT and non-IT equipment. Challenges in this phase related to the existing floor plan/layout, equipment used, and data integration. Due to the budgeting, we have to design the way to install the meters with a reasonable cost. Since CRACs and servers are the major cost of power usage, we should log the power usage from these equipment as a first priority.

After analyzing the electric layout in 2(b), we installed the logger for all the above UDB's. We can log the power usage for server racks, and for corresponding CRACs. The data from IUDB allows us to monitor the power consumption on UPS while the data from NDB allows us to log power consumption for the network room. The data from the logger for MDB shows the power consumption for the whole building where we can use to find the power consumption from other sources (not data center).

We explored the existing equipment and found that some have meters with data logger outputs and some does not. The typical output is RJ-485. The existing ones that have meters with data logger outputs are Schindler meters. The others require the new meter installations.



(a)

(b)

Fig. 2 **a** Floor plan of the main electricity room (MDB) **b** Meter connection layout

We selected the vendors for meter installation. They proposed the choices of meters as well as the specifications. Main criteria for choosing meters are data logger support, and price. The two candidates the vendors proposed that fit our budget were meters from Panasonic and Circutor. Panasonic Eco Power meters have its own programs for data logging and visualization. Thus, we preferred the second one, Circutor (modeled CVM-C10), which was also less expensive and we can write our program to connected to RS-485 and Modbus RTU. The two are common protocol which are supported by most hardware drivers. As a result, we installed the Circutor meters. The meter installation must be done carefully, not to disconnect the running operation. A Labview program was written to read meter data at low-level, and save the data to the database.

We installed the meters on server room 1, which the main server room of the university containing all mail servers and the registration system. The metered were installed for ULC5, ULC6 for both servers' racks and for CRAC3, CRAC4 correspondingly (in Fig. 1). Three months later, all remaining meters were installed

The power meters from Circutor were installed as in Fig. 3a. for every CRAC's and ULC's in the MDB room. The electric cabinets were built for these meters as shown in Fig. 3b. The meter output port is RS-485 and was converted using a serial device server by Advantech, modeled EKI-1522. The LAN cable from EKI was wired to the NOC room where the university network switches are located. The fixed IP was given to the device.

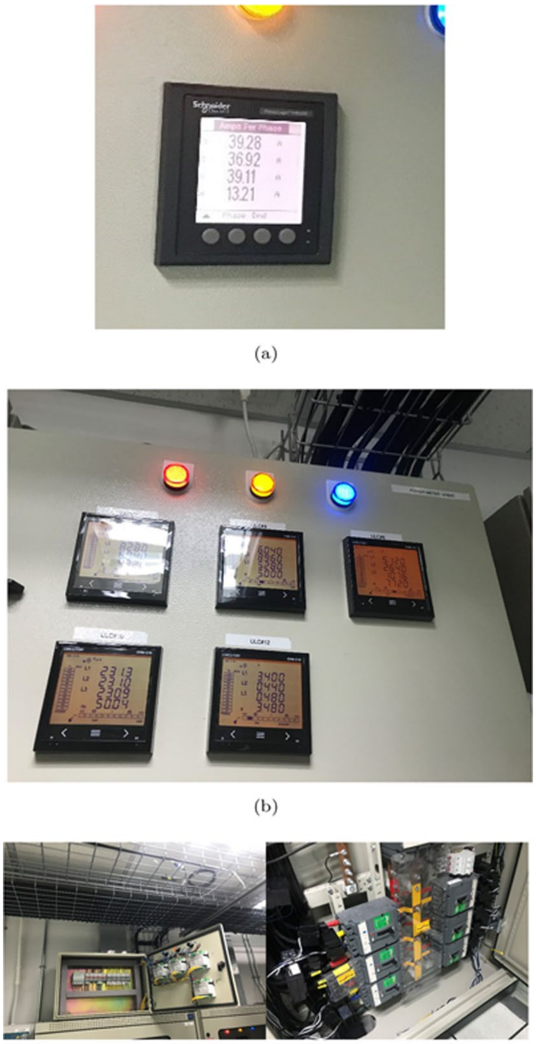
Inside the data center, the Schneider electric control cabinets were installed for IUDB, NDB, ADB, UDB. These meters already have the RJ485 interfaces as shown in Fig. 3a. However, they had no data logger connected to them. Thus, we connected the RS-485 cables to the EKI for a data logger. All the DB meters were used for monitoring the aggregate power consumption to estimate the power consumption of other equipment such as networking and UPS.

After each installation, we measured the error from the meter reading and from the manual clamp. As shown in Fig. 4, the current readings were 39.23A, 36.24A, 39.51A while the reading from the clamps were 39.7A, 36.2A, and 40.6A in 1), 2), 3) respectively. It was found that the errors were about 1A which was acceptable.

As a result, totally 25 m were installed and connected to EKI data logger for collecting all energy usages (according to the color boxes in Fig. 2b).

Data aggregation

Figure 5 presents the overall system for data integration. On the left side, the meter data were collected using the equipment as described above. The data logger in the data center was connected to the data logger server using the TCP/IP. The fixed IP was given to the data logger with the courtesy of OCS. The server was installed MS-SQL 2017 as a database for storing the meter data. Labview 2018 program was implemented to connect the data logger, transform the lower-level data format to the numerical values, and store them in the database. In this paper, the focused portion is shown in the dashed box. The logged data can be further used for visualization and open data purposes. One meter data were kept separately in each database table as in Fig. 6. There are 25 tables for all 25 m.



(a)

(b)

(c)

Fig. 3 **a** Previously installed Schneider meters. **b** Circutor meters. **c** Wiring for the meters

Results

The presented dataset is divided based on three types of installed meters. The first type is the meters for CRAC which measures the air condition units. The second one is the meters for ULC which measure the rack server units. At last, the DB meters measure the aggregate power consumption indicated in the MDB room in the floor plan.

The collection date began on 2 June 2018. At the time of writing the article, the submitted collection has an end date of 18 April 2022. The exact starting dates vary due to the phases of meter installation. There were some periods in the year 2019 and year 2020 when the data logging stopped due to the power outage and downtime of the logging computer. The sampling period is set to 1 min for each meter. This can be set in the data logger program. The total data size is around 21 GB in MS-SQL (2017) backup format. The data was also exported into 25 CSV files, each of which has the varying number of rows and starting dates as shown in Table 1.

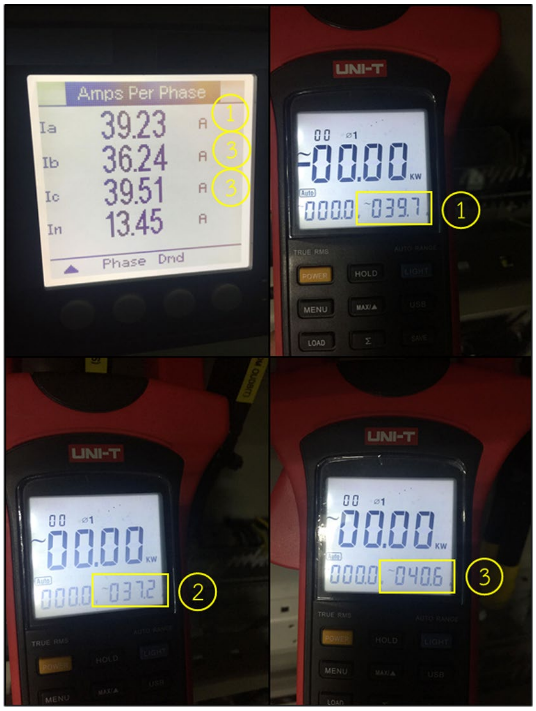


Fig. 4 Comparison of installed meter values and clamp reading

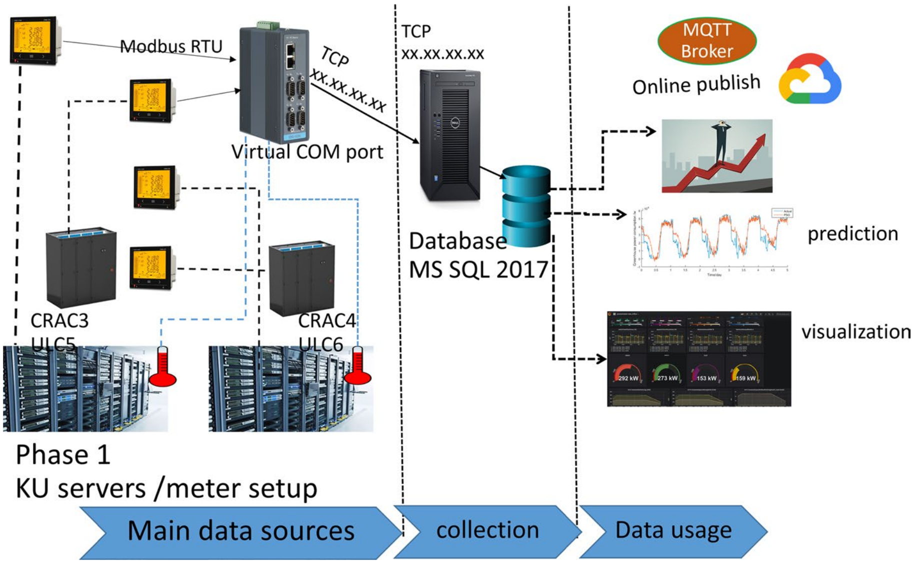
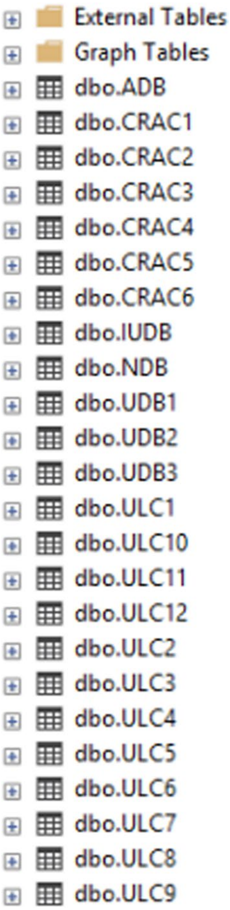


Fig. 5 System and data integration

Each meter data contains 56 fields described in Table 2. In this table, all attributes from the meters were read. The example values are also shown in the table. Figure 7 shows the data rows in the CSV format.

Figures 8a and 9a show example scatter plots for the active three phase for all CRACs and ULCs in our dataset while Fig. 8b and 9b show the box plots of the data.



External Tables
Graph Tables
dbo.ADB
dbo.CRAC1
dbo.CRAC2
dbo.CRAC3
dbo.CRAC4
dbo.CRAC5
dbo.CRAC6
dbo.IUDB
dbo.NDB
dbo.UDB1
dbo.UDB2
dbo.UDB3
dbo.ULC1
dbo.ULC10
dbo.ULC11
dbo.ULC12
dbo.ULC2
dbo.ULC3
dbo.ULC4
dbo.ULC5
dbo.ULC6
dbo.ULC7
dbo.ULC8
dbo.ULC9

Fig. 6 Logged data tables in the database

The average values for CRAC3-CRAC4 are 10,959.92 and 6,665.48 while the average values for ULC5-ULC6, are 6108.42, and 6531.58 respectively.

Data processing and analysis

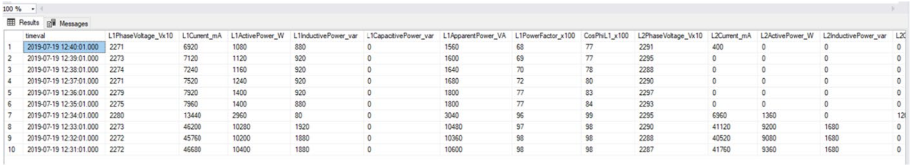
Data cleansing

In this section, we demonstrate the usage of the dataset. As an example, the first 3-month data from CRAC3, CRAC4 and ULC5, ULC6 were extracted as CSV files. Jupyter Notebook, python language and relevant libraries such as pandas, numpy, sklearn, matplotlib, pmdarima, joblib were used to preprocess and visualize the data. The data received may be temporarily disconnected due to network connection. The sensor values may be outliners.

1. removing extreme values: to remove outliners by inspecting the negative values and non increasing values.
2. filling missing data: missing data due to network connections needs to be filled.
3. removing unneeded columns: columns that do not need for model creation.

Table 1 Total data sizes and date ranges for the dataset

Dataset	Total rows	Sampling period	Starting date	Enddate
CRAC1	1570231	1 min	2018-08-09	2022-04-12
CRAC2	1578407	1 min	2018-08-09	2022-04-18
CRAC3	1685982	1 min	2018-07-10	2022-04-12
CRAC4	1685982	1 min	2018-07-10	2022-04-12
CRAC5	1578401	1 min	2018-08-09	2022-04-12
CRAC6	1578402	1 min	2018-08-09	2022-04-12
ULC1	1578403	1 min	2018-08-09	2022-04-12
ULC2	1578398	1 min	2018-08-09	2022-04-12
ULC3	1578406	1 min	2018-08-09	2022-04-12
ULC4	1578398	1 min	2018-08-09	2022-04-12
ULC5	1686001	1 min	2018-08-09	2022-04-12
ULC6	1686009	1 min	2018-08-09	2022-04-12
ULC7	1578414	1 min	2018-08-09	2022-04-12
ULC8	1578408	1 min	2018-08-09	2022-04-12
ULC9	1578413	1 min	2018-08-09	2022-04-12
ULC10	1578409	1 min	2018-08-09	2022-04-12
ULC11	1589936	1 min	2018-08-09	2022-04-26
ULC12	1578409	1 min	2018-08-09	2022-04-12
ADB	1570252	1 min	2018-08-09	2022-04-12
NDB	570242	1 min	2018-08-09	2022-04-12
IUDB	1570221	1 min	2018-08-09	2022-04-12
UDB1	1578444	1 min	2018-08-09	2022-04-12
UDB2	1578442	1 min	2018-08-09	2022-04-12
UDB3	1578455	1 min	2018-08-09	2022-04-12



	timeval	L1PhaseVoltage_Vx10	L1Current_mA	L1ActivePower_W	L1InductivePower_var	L1CapacitivePower_var	L1ApparentPower_VA	L1PowerFactor_x100	CosPhiL1_x100	L2PhaseVoltage_Vx10	L2Current_mA	L2ActivePower_W	L2InductivePower_var	L2C
1	2019-07-19 12:40:01.000	2271	6920	1080	880	0	1560	68	77	2291	400	0	0	0
2	2019-07-19 12:39:01.000	2273	7120	1120	920	0	1600	69	77	2295	0	0	0	0
3	2019-07-19 12:38:01.000	2274	7240	1160	920	0	1640	70	78	2288	0	0	0	0
4	2019-07-19 12:37:01.000	2271	7520	1240	920	0	1680	72	80	2290	0	0	0	0
5	2019-07-19 12:36:01.000	2279	7920	1400	920	0	1800	77	83	2297	0	0	0	0
6	2019-07-19 12:35:01.000	2275	7960	1400	880	0	1800	77	84	2293	0	0	0	0
7	2019-07-19 12:34:01.000	2280	13440	2960	80	0	3040	96	99	2295	6960	1360	0	12
8	2019-07-19 12:33:01.000	2273	46200	10280	1920	0	10480	97	98	2290	41120	9200	1680	0
9	2019-07-19 12:32:01.000	2272	45760	10200	1880	0	10360	98	98	2288	40520	9080	1680	0
10	2019-07-19 12:31:01.000	2272	46680	10400	1880	0	10600	98	98	2287	41760	9360	1680	0

Fig. 7 Example data rows in CSV

The first two steps were required since there were always some error data due to some bad reading sensors. The missing data were due to some power outage periods and the intermittent disconnection of the logging server. The extreme values must be observed for each attribute to find out the maximum and minimum limits. The scheme for filling in missing values may be using mean, summation, and filling with zero, depending on the attributes. Different methods can affect the correctness of the models. Since it is time-series data, history data is used for model construction. Filling with zero or filling with mean can lead to different prediction model. This also depends on how many missing values are continuously. Imputing more values can lead to noises in data. Thus, it should be done with care (Hyndman and Athanasopoulos 2018).

Depending on the goal and the method, the third step may remove some columns. For a demonstration, we are interested in the power consumption attribute. The

Table 2 Sample fields for meter data collection

Field name	Example value
timeval	10/7/2018 08:01
L1 Phase voltage	2285
L1 Current	33160
L1 Active Power (W)	6960
L1 Inductive Power	2800
L1 Capacitive Power	0
L1 Apparent Power	7560
L1 Power Factor (×100)	91
Cos ϕ L1 (×100)	92
L2 Phase voltage	2310
L2 Current	28240
L2 Active Power (W)	6160
L2 Inductive Power	1960
L2 Capacitive Power	0
L2 Apparent Power	6520
L2 Power Factor (×100)	94
Cos ϕ L2 (×100)	95
L3 Phase voltage	2305
L3 Current	27560
L3 Active Power (W)	5840
L3 Inductive Power	2360
L3 Capacitive Power	0
L3 Apparent Power	6320
L3Power Factor (×100)	91
Cos ϕ L3(x100)	92
Active Three-phase Power	19000
Inductive Three-phase power	7160
Capacitive Three-phase Power	0
Apparent three-phase power	20400
Three-phase Power Factor	93
Three-phase Cos ϕ	93
L1 Frequency (x100)	5007
L1-L2 Voltage	3971
L2-L3 Voltage	4000
L3-L1 Voltage	3978
Neutral Current N(mA)	7720
Inductive Three-phase power	6880
Capacitive Three-phase power	0
Apparent three-phase power	20160
Three-phase Power Factor	93
Three-phase Cos ϕ	93
L1 Frequency (x100)	5004
L1-L2 Voltage	3934
L2-L3 Voltage	3963
L3-L1 Voltage	3945
Neutral Current N(mA)	7640
L1 voltage % THD	20
L2 voltage % THD	16
L3 voltage % THD	19

Table 2 (continued)

Field name	Example value
L1 current % THD	125
L2 current % THD	127
L3 current % THD	133
Maximum demand kW III	20160
Maximum demand kVA III	21360
Maximum demand I AVG	31080
Maximum demand I L1	33440
Maximum demand I L2	30240
Maximum demand I L3	29560
Consumed active energy kW)	622
Consumed active energy (W)	962
Consumed inductive reactive energy (kvarhL)	188
Consumed inductive reactive energy (varhL)	279
Consumed capacitive reactive energy (kvarhC)	0
Consumed capacitive reactive energy (varhC)	11
Consumed apparent energy (kVAh)	654
Consumed apparent energy (VAh)	43
Consumed CO2 emissions	0

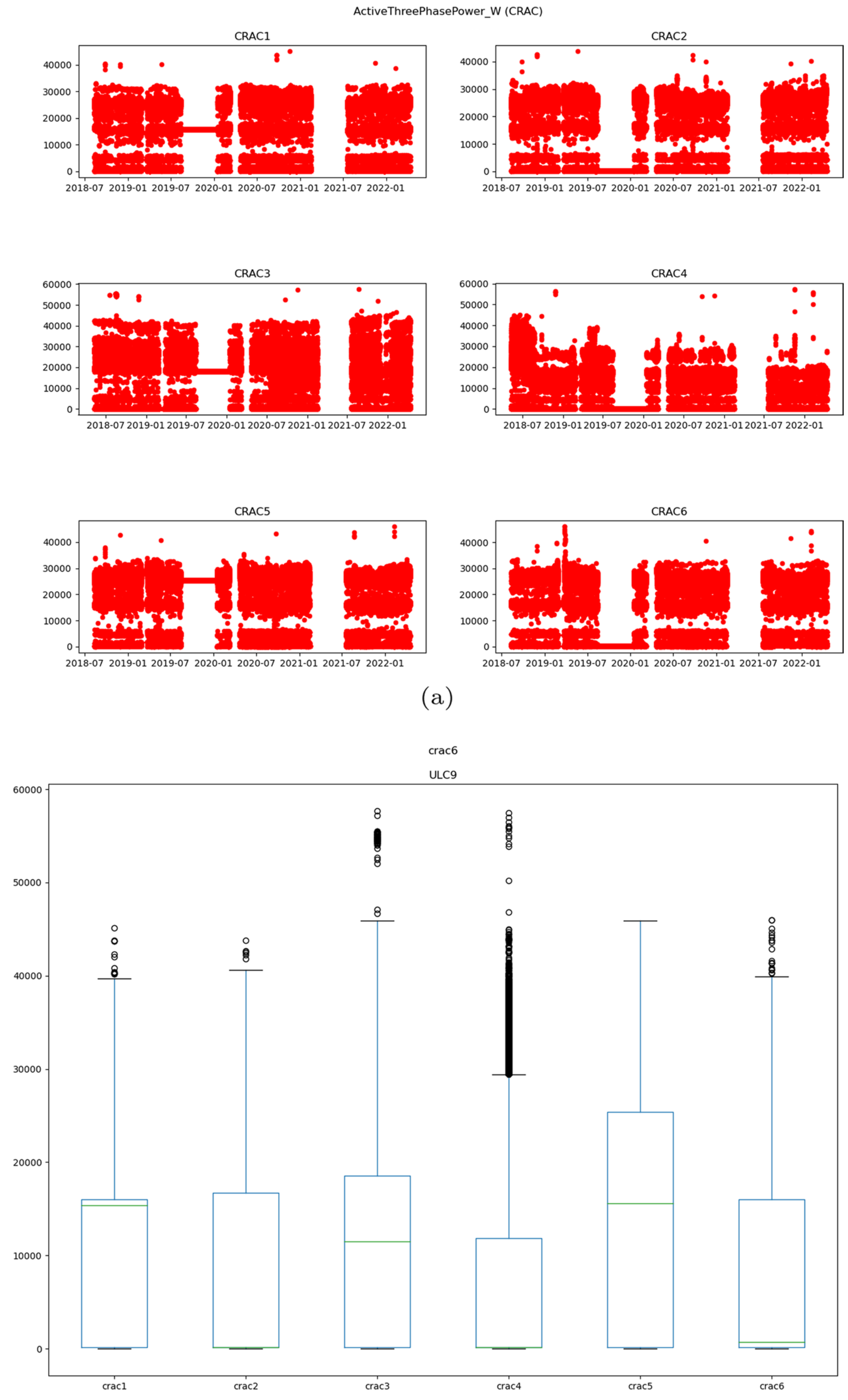
column "Consumed active energy kW" was chosen as a target column. This column is an accumulated value computed from "Active three phase kW".

Figure 10 shows the example data for 3 months of such values for CRAC3, CRAC4 and ULC5,ULC6 in Fig. 10a and b respectively. It can be seen that there are some missing values and some error values. Therefore, data cleansing can be done. From the visualization, we set the maximum value for the y-axis as 30,000 and the minimum value for y as 0. We checked the non-increasing values for the y-axis. When we found the values over the maximum values and the values that were not non-increasing ones, the associated rows were dropped. In Fig. 11, the plotted of data after dropping values are shown in Fig. 11a, b for CRAC3, CRAC4 and ULC5, ULC6 respectively.

The total number of wrong values found is displayed in Table 3 Column "less" contains the number of values less than equal to zero, Column "exceed" contains the number of values more than 30,000, and Column "non-decrease" are the number of values that are non-decreasing. Figure 12 presents the resulting data after using filling with different approaches for CRAC3 and CRAC4. In Fig. 12a, b, *means* and *bfill* (backward fill) (Pandas 2024) approaches were used to fill the dropped rows respectively. Figure 13 shows the resulting data for ULC5, ULC6 with the same filling approaches, i.e., Fig. 13a, b, for *means* and *bfill* approaches respectively.

Data analysis

After performing, the data cleansing as in the previous section, the data can be used to construct the prediction models. Since it is time-series data, we consider three types of models are presented: regression, autoregression, and ARIMA. In each type of model, specific parameters are explored. In regression models, one have to decide the attributes used. The correlations of attributes have to be studied. For autoregression, the lag window size must be chosen. Finally, ARIMA combines autoregression, moving average



(a)

(b)

Fig. 8 Active three phase power for all CRACs in the dataset **a** scatter plot **b** box plot

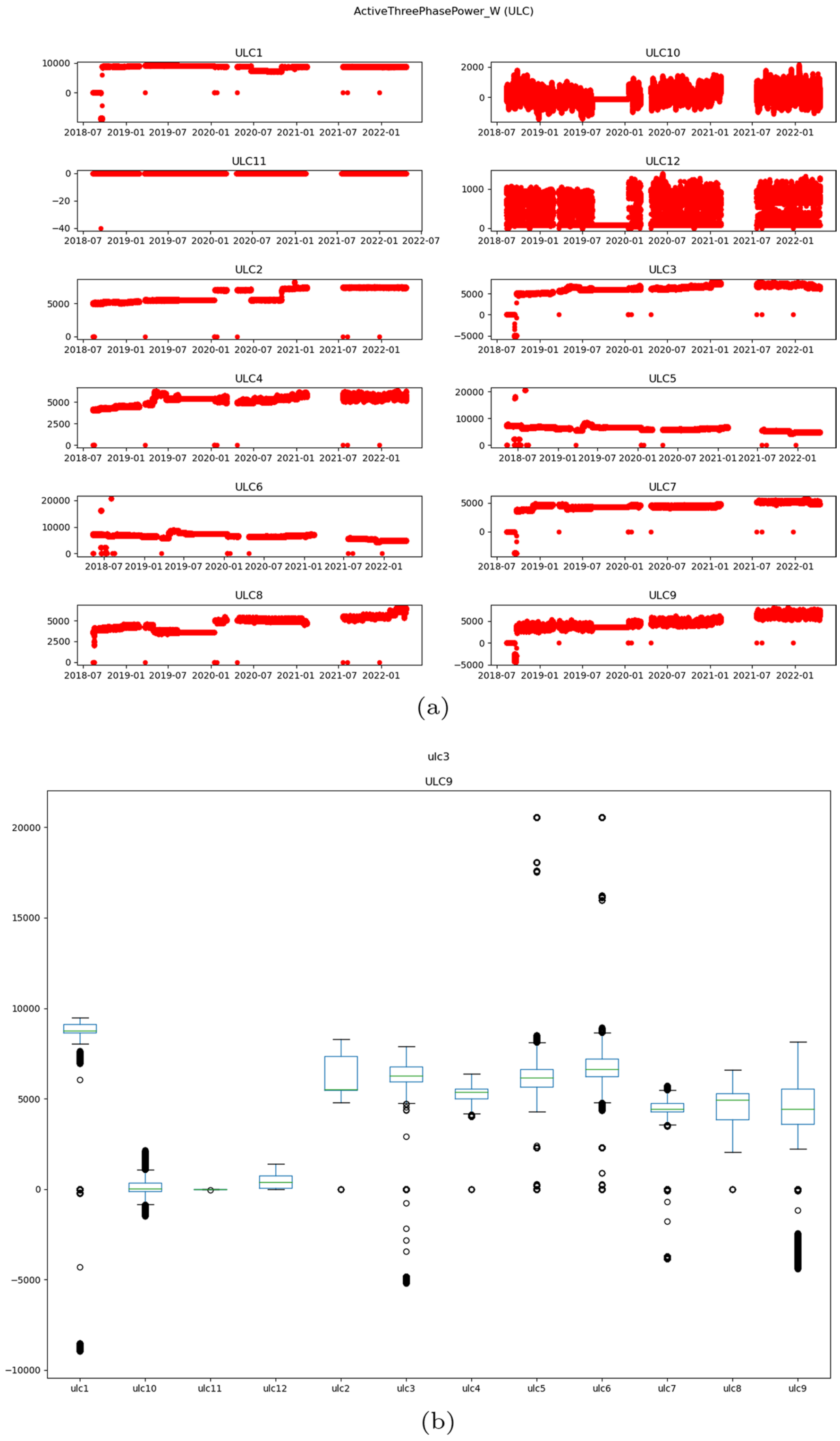


Fig. 9 Active three phase power for all ULC in the dataset **a** scatter plot **b** box plot

Table 3 Total missing values and wrong values

Data	Total rows	Less	Exceed	Non-decrease	Cleaned line
CRAC3	393,108	54	4	1887	96,461
CRAC4	393,108	39	59	1511	96,791
ULC5	393,108	1226	246	2067	95,755
ULC6	393,108	1185	54	1906	96,363

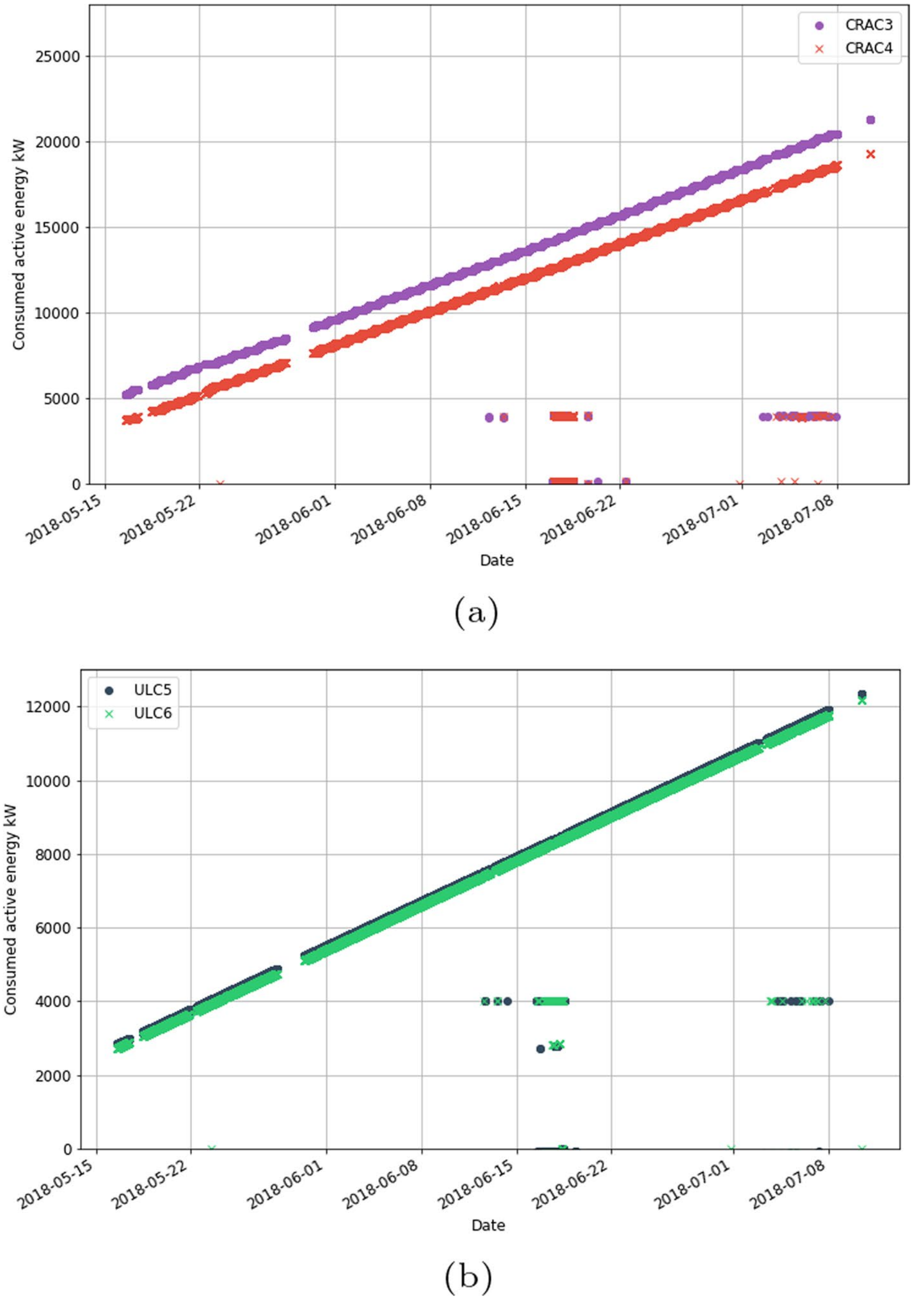
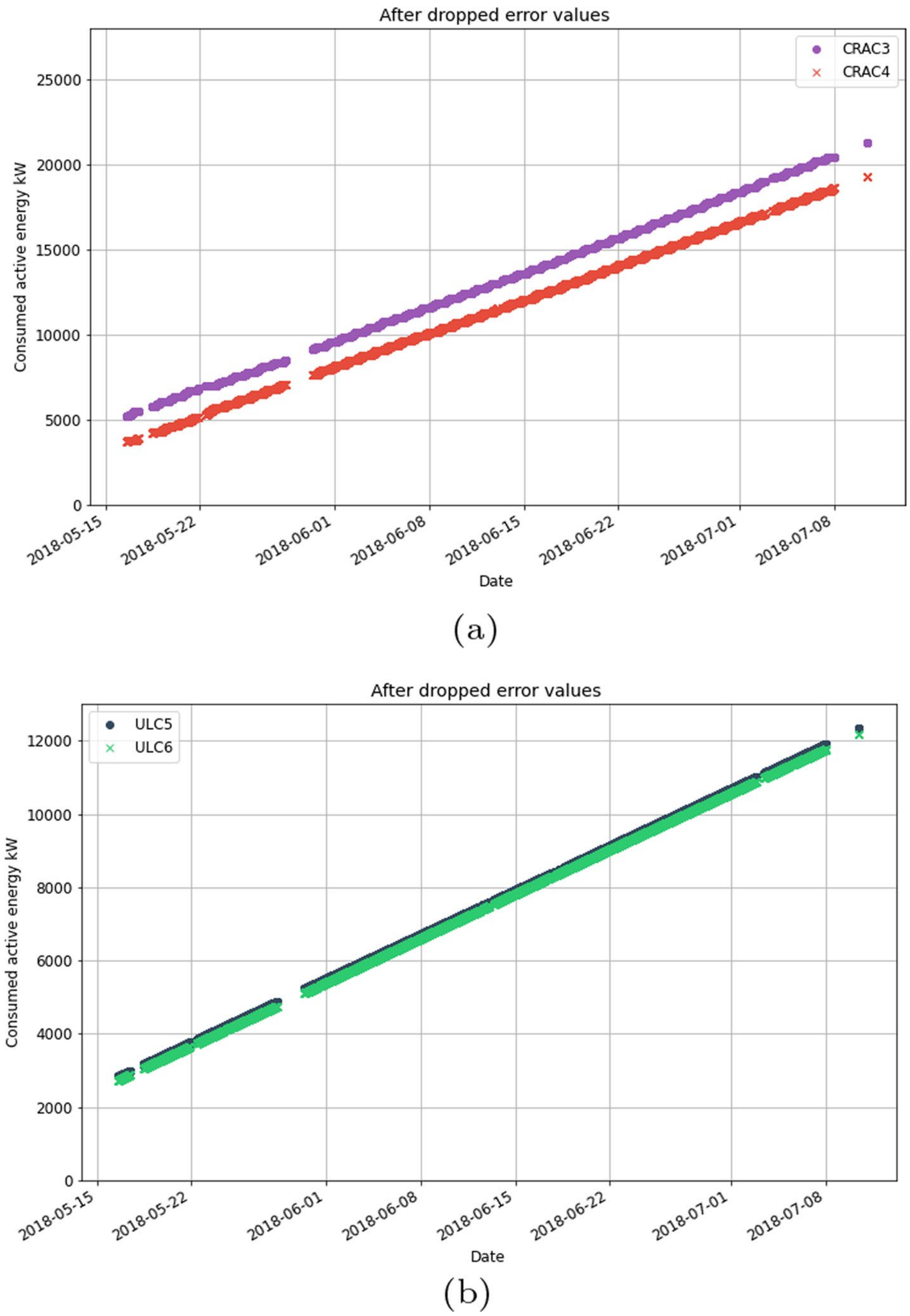


Fig. 10 Consumed active power **a** CRAC3 and CRAC4 **b** ULC5 and ULC6



(a)

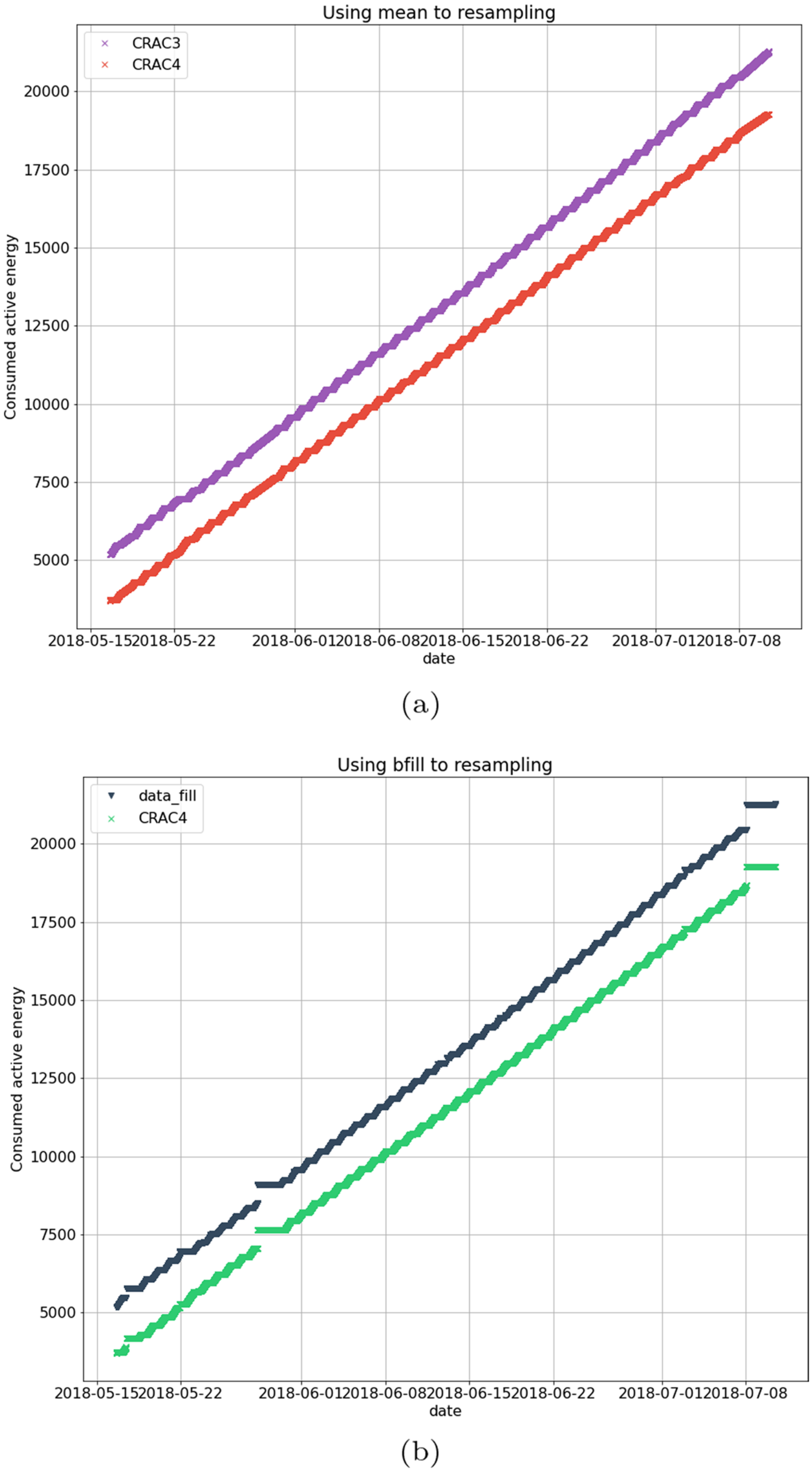
(b)

Fig. 11 Consumed active power (drop values) **a** CRAC3 and CRAC4 **b** ULC5 and ULC6

windows, and degree difference. The model is more complicated. More parameters need to be explored. In the following, we demonstrate the use of these models against our dataset.

Multivariable regression

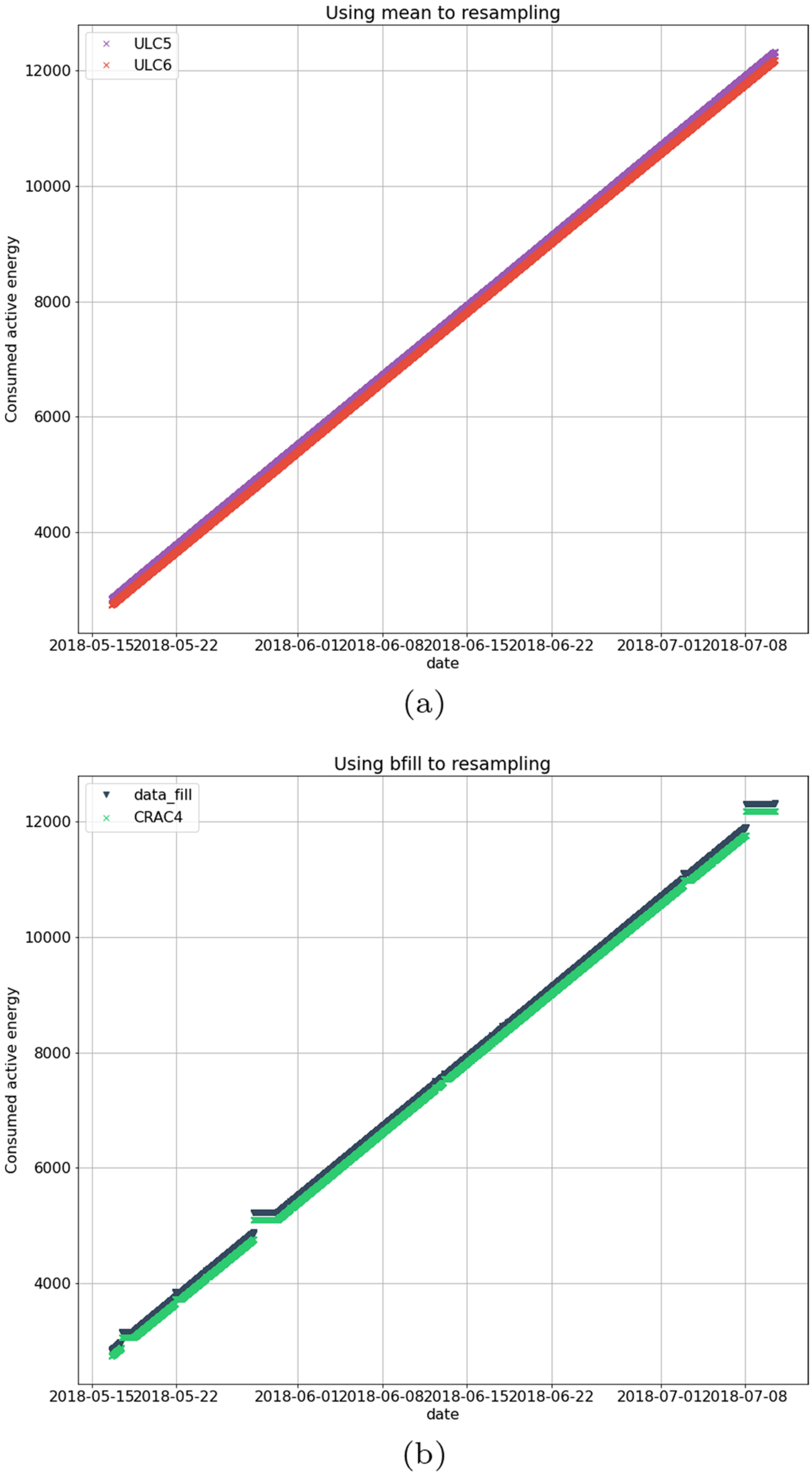
Using the multivariable regression approach (sklearn 2024), several variables were considered whether there are relationship to the target variable. Since there were many



(a)

(b)

Fig. 12 CRAC3 and CRAC4 consumed active power **a** mean **b** bfill



(a)

(b)

Fig. 13 ULC5 and ULC6 consumed active power **a** mean **b** bfill

attributes, we explored important variables using co-relations. We use *corr()* function in pandas to calculate. In this function, there are 3 types of co-relations: Pearson, Kendall, and Spearman (Pandas 2024).

Example correlation plots between Consumed active energy (kW) and consumed apparent energy(kVAh), consumed inductive reactive energy (kvarhL), consumed capacitive reactive energy (kvarhC) of CRAC3 are shown in Fig. 14. After calculating the co-relations, the ranking of the co-relations between "Consumed active energy" and other fields. Assume we picked the top 3 attributes with the highest co-relation values.

Using the linear regression model, the prediction for each CRAC3 and CRAC4 is shown in Fig. 15. Root means square error (RMSE) values are 0.95 and 0.94 respectively. In Fig. 16, RMSE values are 0.98 and 0.99 for ULC5 and ULC6 respectively. Figure 15 shows some noises in prediction values since the attribute values do not get cleaned except consumed active energy (kW) values.

Auto-regression

Auto-regression is a persistent model which predicts the point at $t + 1$ from the previous time point t after that the new data point is appended and the process continues. If it is based on one previous point, lag = 1. Since this model only requires previous time points, only one field was used to create model, "consumed active energy". The other fields were removed. Figure 17 presents RMSE values for CRAC3, CRAC4, ULC5, and ULC6 for the same data used in the previous subsection. We shifted the data by one point time unit in order to create training data for lag = 1. The total data points were 45,339 with the mean value 8175 kW. We set 67% of them to be trained data and the remaining was test data. This resulted in 48,887 training rows and 24,080 testing rows. The reported RMSE is based on the test data. The RMSE is small compared to using the linear regression approach.

We implemented the auto-regression model using *statsmodel* in the python library (statsmodels 2024). The method automatically finds proper lag values during fitting. We used the same portion of training and testing. The library finds the number of lags and coefficients. The RMSE results are shown in Fig. 18. All of the shown cases have lag = 50 after finishing training. The RMSE values were higher compared to Fig. 17 for all the four meters.

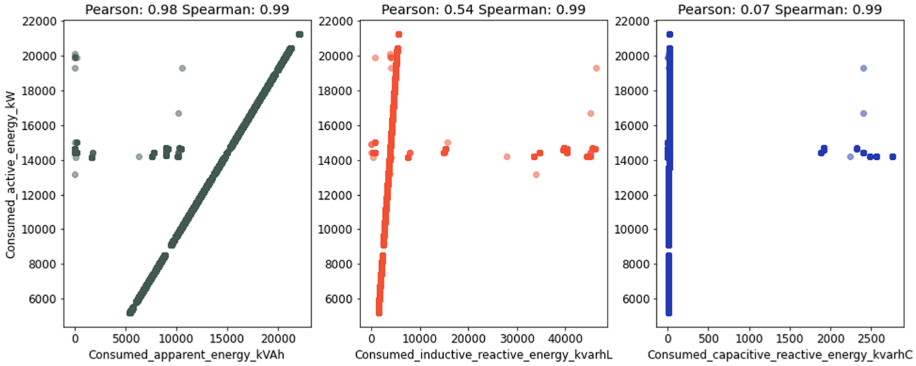
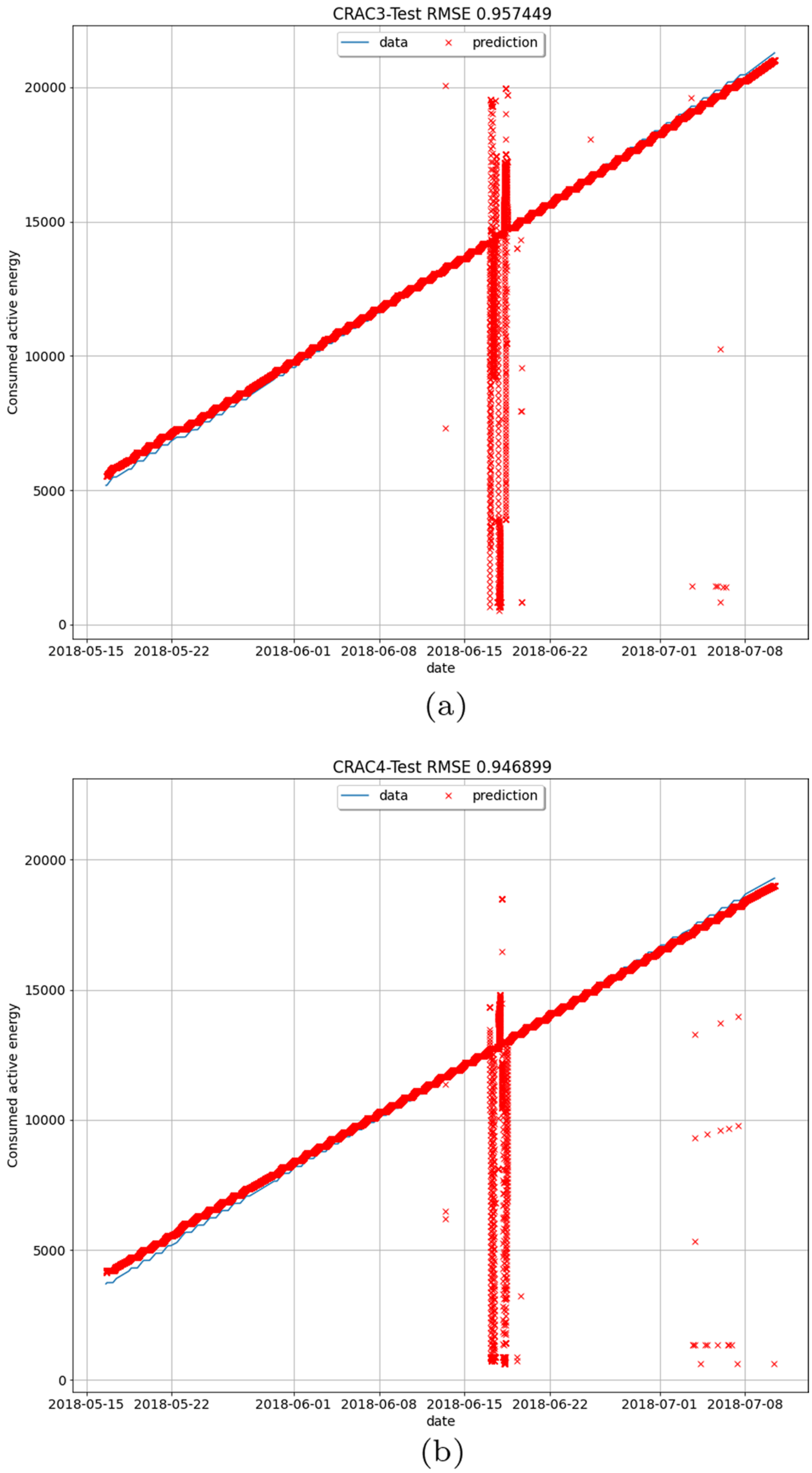


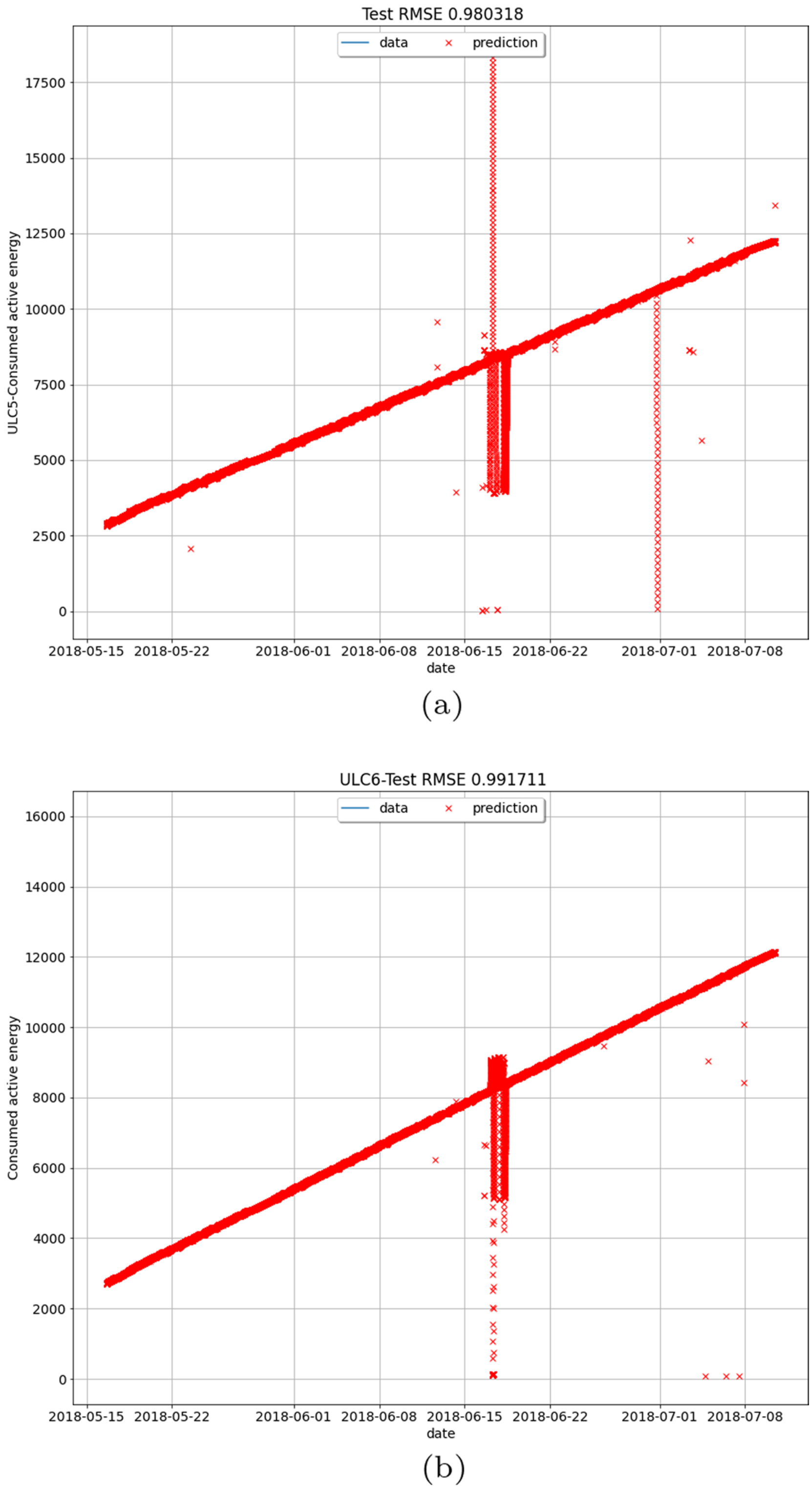
Fig. 14 CRAC3 Spearman correlation



(a)

(b)

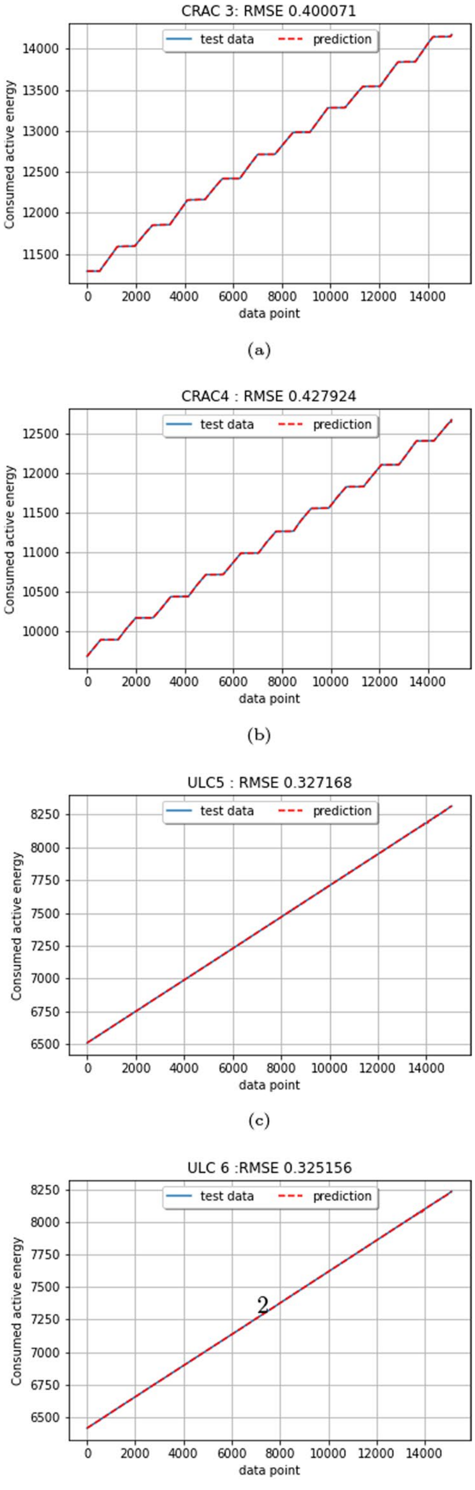
Fig. 15 Consumed active power regression **a** CRAC3 **b** CRAC4



(a)

(b)

Fig. 16 Consumed active power regression **a** ULC5 **b** ULC6



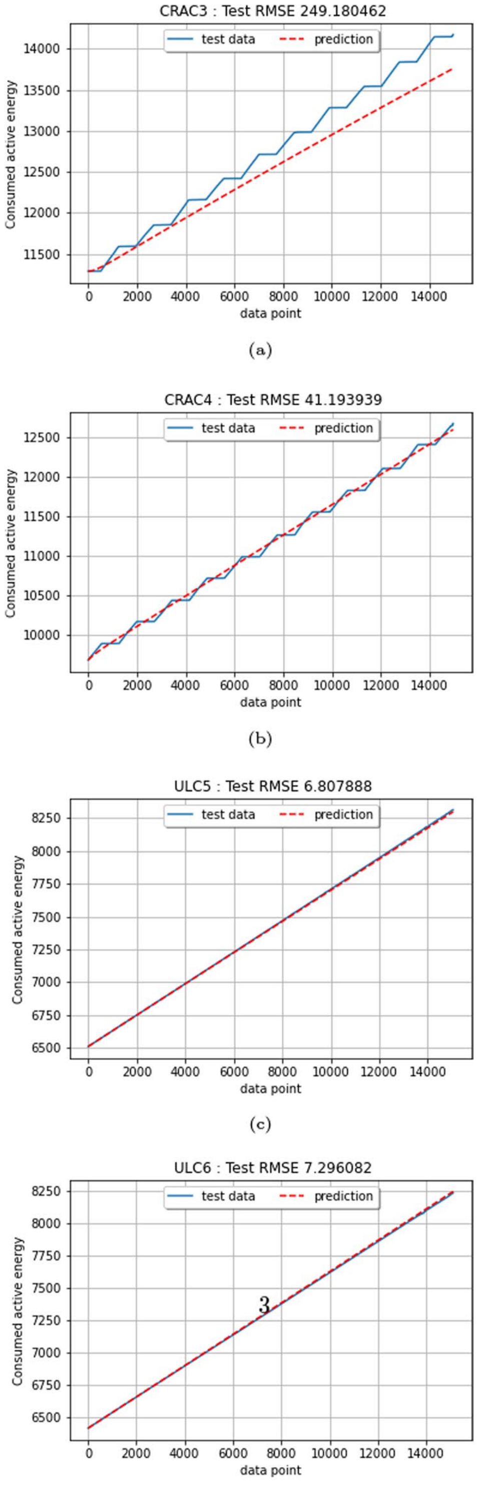
(a)

(b)

(c)

(d)

Fig. 17 Auto-regression (I) **a** CRAC3 **b** CRAC4 **c** ULC5 **d** ULC6



(a)

(b)

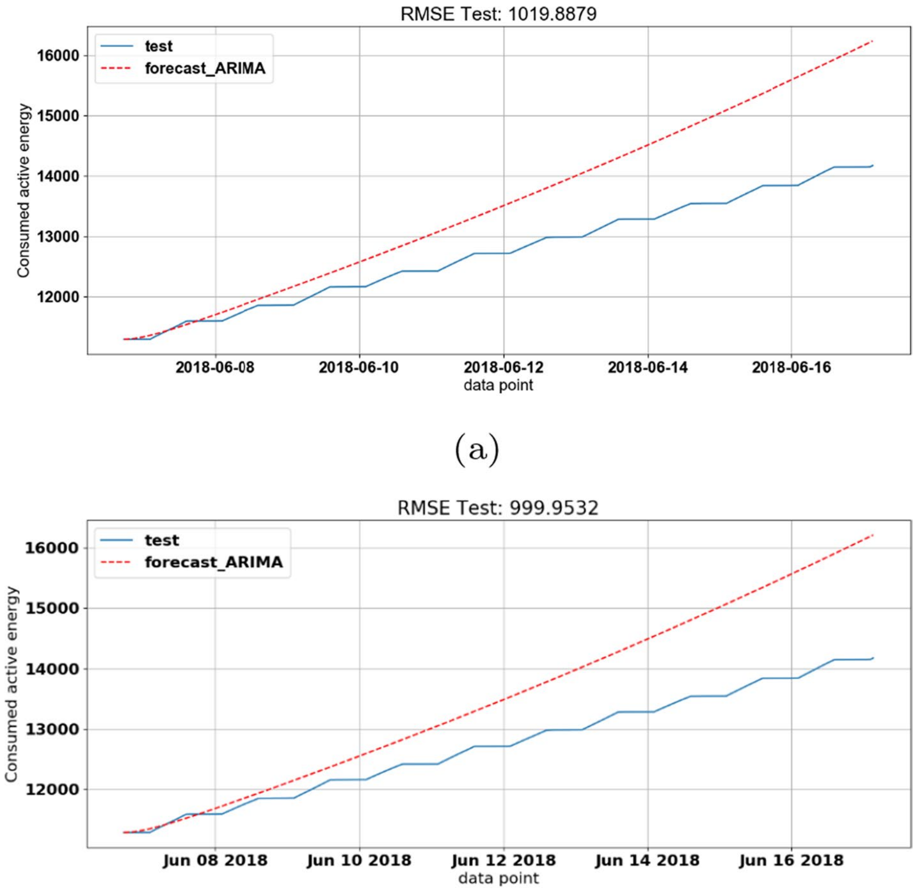
(c)

(d)

Fig. 18 Auto-regression (II) **a** CRAC3 **b** CRAC4 **c** ULC5 **d** ULC6

ARIMA

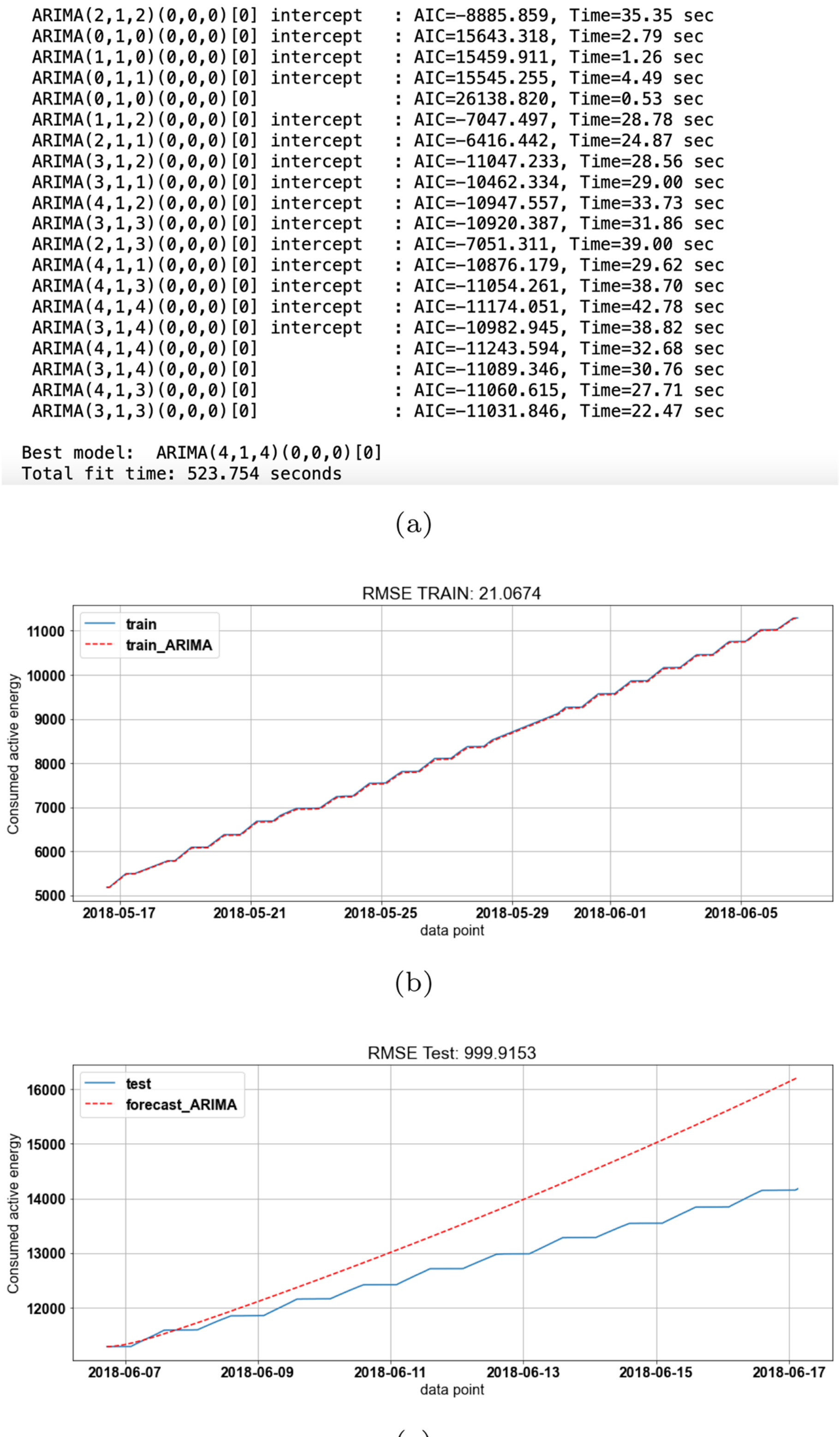
Auto-Regressive Integrated Moving Averages (ARIMA) (Hyndman and Athanasopoulos 2020) can be used with stationary data. ARIMA has 3 important parameters p, d, and q. p is the lag time, q is the number of moving average terms which is a lagged value forecasting error. For example, if $q = 5$, for $y(t)$, we have $e(t-1), e(t-2), \ldots e(t-5)$, where $e(i)$ represents the difference between the moving average at i^{th} instance and at the moving average at t. At last, d is the order of difference, e.g., $d = 1$ is the first order difference. For constructing the ARIMA model, p and q values need to be estimated. The auto-correlation function (ACF) and partial auto-correlation function (PACF) were measured (Hyndman and Athanasopoulos 2020). Varying p, q, and d incurs different RMSE values. Figure 19 shows the case when using $p = 2, d = 1, q = 2$ and $p = 3, d = 1, q = 3$ for CRAC3. When performing the grid search, the best p, d, and q were found. Figure 20a presents the parameter spaces in the grid search for CRAC3 while Fig. 20c presents the RMSE when using the best parameter $p = 4, d = 1$, and $q = 4$ with the training set. For the purpose of demonstration only, the number of trained data rows is 30,409 and the number of test data rows is 14,979. The trained data achieves RMSE, 21.06. The test data has a larger



(a)

(b)

Fig. 19 ARIMA (CRAC3) **a** $p = 2, d = 1, q = 2$ **b** $p = 3, d = 1, q = 3$



```
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=-8885.859, Time=35.35 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=15643.318, Time=2.79 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=15459.911, Time=1.26 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=15545.255, Time=4.49 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=26138.820, Time=0.53 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=-7047.497, Time=28.78 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=-6416.442, Time=24.87 sec
 ARIMA(3,1,2)(0,0,0)[0] intercept   : AIC=-11047.233, Time=28.56 sec
 ARIMA(3,1,1)(0,0,0)[0] intercept   : AIC=-10462.334, Time=29.00 sec
 ARIMA(4,1,2)(0,0,0)[0] intercept   : AIC=-10947.557, Time=33.73 sec
 ARIMA(3,1,3)(0,0,0)[0] intercept   : AIC=-10920.387, Time=31.86 sec
 ARIMA(2,1,3)(0,0,0)[0] intercept   : AIC=-7051.311, Time=39.00 sec
 ARIMA(4,1,1)(0,0,0)[0] intercept   : AIC=-10876.179, Time=29.62 sec
 ARIMA(4,1,3)(0,0,0)[0] intercept   : AIC=-11054.261, Time=38.70 sec
 ARIMA(4,1,4)(0,0,0)[0] intercept   : AIC=-11174.051, Time=42.78 sec
 ARIMA(3,1,4)(0,0,0)[0] intercept   : AIC=-10982.945, Time=38.82 sec
 ARIMA(4,1,4)(0,0,0)[0]             : AIC=-11243.594, Time=32.68 sec
 ARIMA(3,1,4)(0,0,0)[0]             : AIC=-11089.346, Time=30.76 sec
 ARIMA(4,1,3)(0,0,0)[0]             : AIC=-11060.615, Time=27.71 sec
 ARIMA(3,1,3)(0,0,0)[0]             : AIC=-11031.846, Time=22.47 sec

Best model:  ARIMA(4,1,4)(0,0,0)[0]
Total fit time: 523.754 seconds
```

(a)

(b)

(c)

Fig. 20 ARIMA (CRAC3) **a** Grid search result $p = 4, d = 1, q = 4$ **b** train **c** test

RMSE, 999.91. The testing score is worse than the training score since the training set may not be generalized enough.

Discussion and implications

Table 4 summarizes accuracy for all methods for the sample dataset. It is seen that the simple method, multivariate linear regression, is simplest and effective. Multivariate regression considers more than one variable which have correlations. The linear regression is suitable for this type of data as we can see later in the visualization section. For autoregression, suitable lag time must be selected. Lag time shows a dependency window for the current time value. One might utilize grid search to find suitable lag time. In this example, few lags are better. When combining with multivariable, it would be more complex to explore both lag time and co-related variables. Similarly, with ARIMA, proper parameters, values of p, d, q must be explored. More data are needed since they can exhibit seasonal nature. In the above example, only 3-month data cannot demonstrate seasonal cycle very well. Longer time period data may be needed.

However, this test dataset is only for 3 months. It is opened that the methods can be trialed against the whole dataset, i.e., for other meters and longer period. The other attributes can also be considered along with other prediction models such as LSTM.

Visualization example

We constructed the sample visualization using GrafanaLab (2024) which shows system statistics, as well as CRAC, and ULC meter values in the same dashboard. The live connection to MS-SQL database was used.

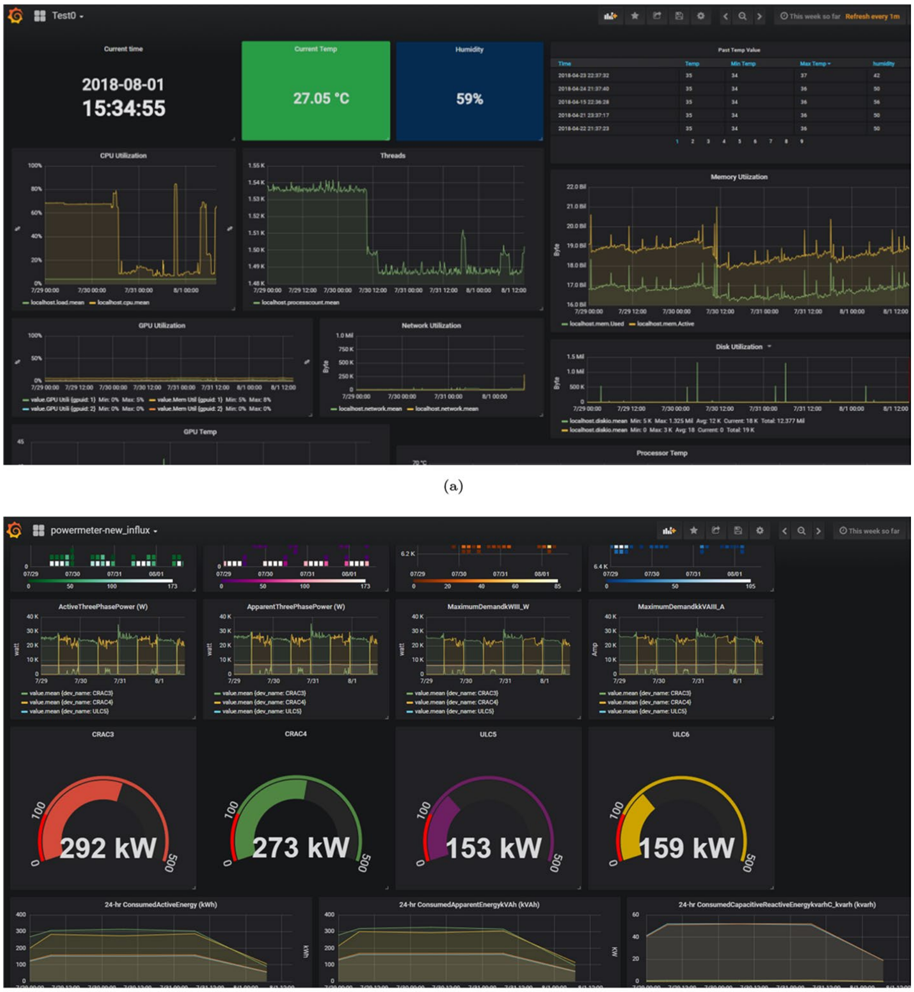
Figure 21b presents the values from MS-SQL database CRAC3, CRAC4, ULC5, ULC6, to show the three phase active power, consumed energy, L1,L2, and L3 active power, etc. during that period. Figure 23 focuses on the three phase active power widgets showing max,min, average values of the four meters.

When consider the integration of the server workload data, Fig. 21a shows the example of logging of power consumption, memory usage, network, GPU, etc. These logging were done using the scripts running as background processes on a Linux system which recorded data to InfluxDB influxdata (2024). Grafana, then, connected to influxDB for the time-series data display.

For future analysis, the dashboard in Fig. 22 shows the year plots of accumulated powers for all the meters in Fig. 2b where the trend line is shown. We analyze the activities on the period with high power usage values, for example during the end of

Table 4 Prediction score of various methods (RMSE)

Data	Multivar Linear	AutoR Lag = 1	AutoR Lag = 50	ARIMA $p = 3, d = 1, q = 3$	ARIMA $p = 4, d = 1, q = 4$
CRAC3	0.94	0.40	249.18	1,236.18	999.95
CRAC4	0.95	0.42	41.19	1,617.19	1,633.50
ULC5	0.98	0.32	6.80	770.15	770.24
ULC6	0.99	0.32	7.29	793.61	793.14



(a)

(b)

Fig. 21 Grafana visualization. **a** Computer loads **b** Energy monitoring

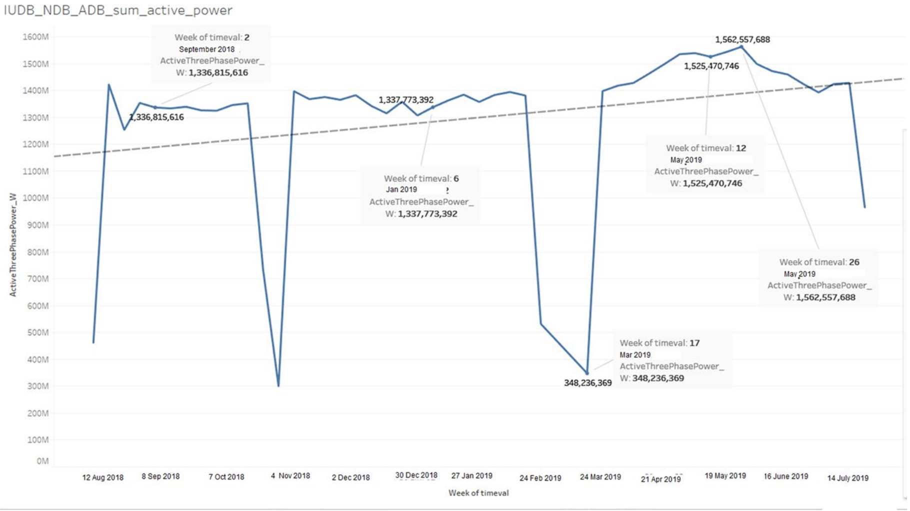


Fig. 22 Overall power consumption analysis and trend

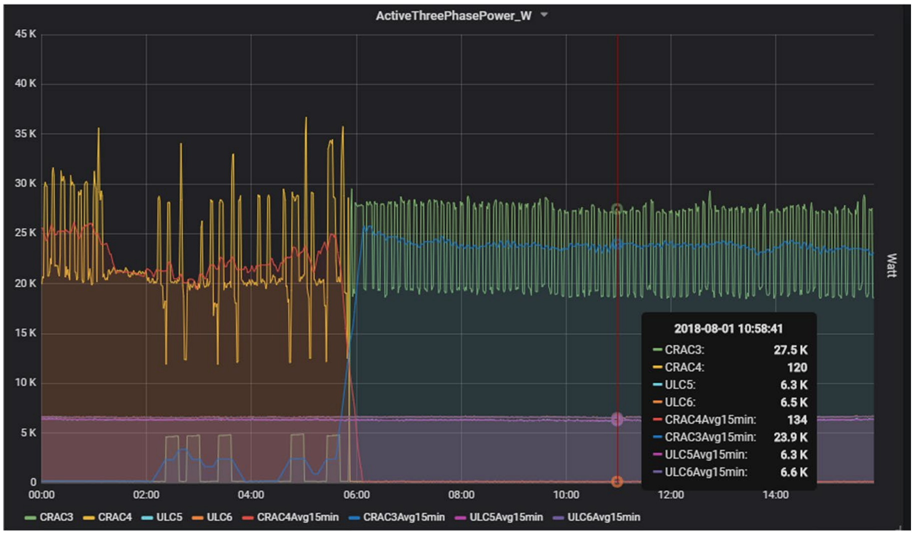


Fig. 23 Grafana visualization. Zoomed three phase active power

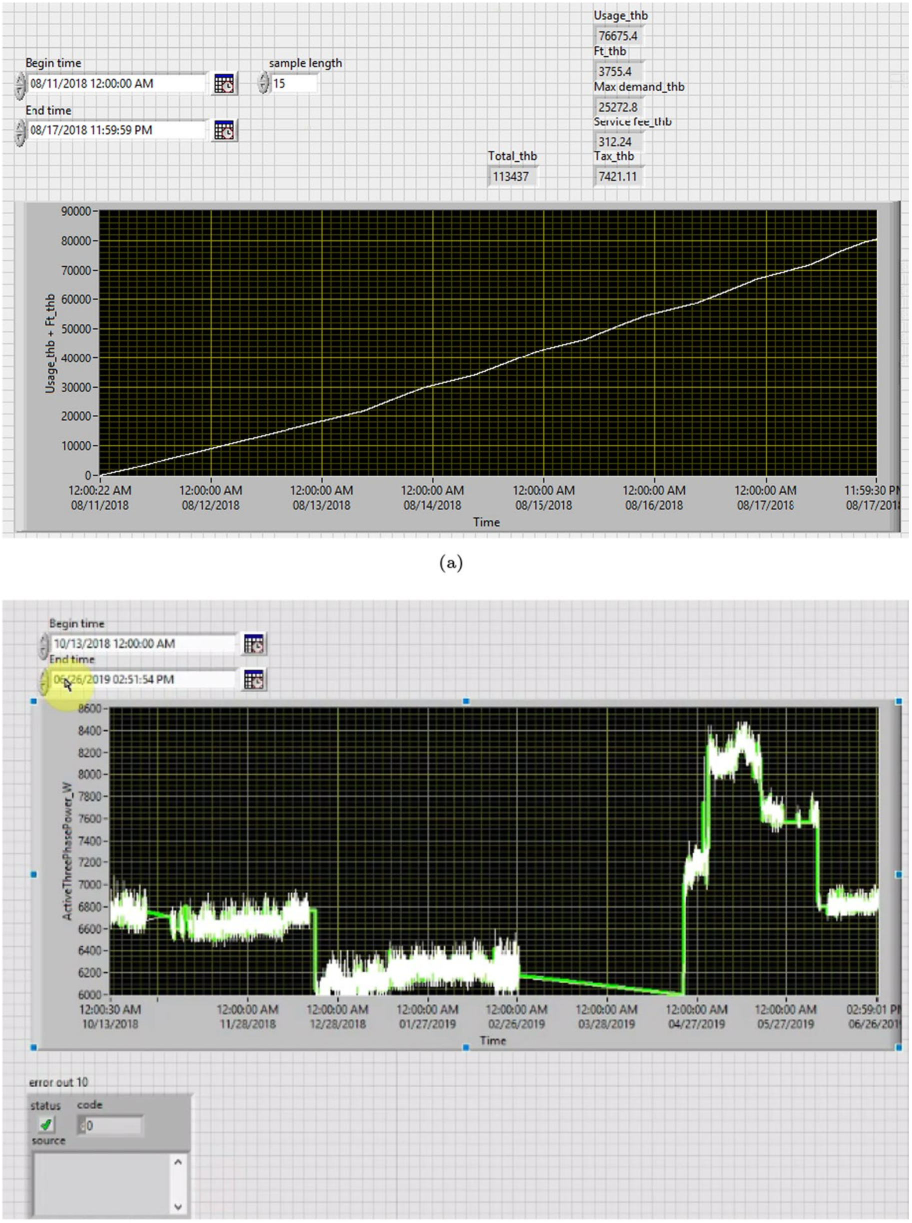
semester period and during the summer period. The trend line suggests the increment of the power usage.

From the visualization, the recommendations for energy saving can be reducing the workloads during the summer period. This may not be directly possible during the seasonal activity in the university. Based on the billing policy PEA (2024) in Thailand, the power cost is differed based on weekdays and weekends, varying day and night. Other saving strategies may be the transfer of high workload to weekends or overnights, etc.

Based on the three-phase active power and the billing policy, we construct the estimation of the billing cost as in Fig. 24a. Figure 24b expresses the moving average of three-phase active power showing the peak consumption day. The visualization can be used for future investigation of the peak consumption and to analyze the causes. In other way, the prediction trend can suggest the billing cost which help future budget planning.

From Loeffler (2016), the 10 most popular way to minimize energy in a data center is to (1) turn off idle equipment or idle server (low utilization : below 15%) (2) use virtualization (3) use blade servers and gather storage (4) turn on power management feature of CPU to change frequency properly according to the utilization (5) use IT equipment with efficient power supply, and chain the IT equipment from electric sources (6) use efficient UPS and PDU (7) use power distribution 208V/230V (8) use good cooling system, use hot aisle and cold aisle (9) measure the energy usage in the data center from output of UPS and calculate PUE (10) prioritize the optimization of energy usage and look for the equipment that is not used.

In our scenario, we cannot move the rack nor upgrade the hardware. The possible strategy is workload management. For example, to delay running workload during the night can save the power cost. Or adjusting the temperature setting for CRAC to save power consumption.



Begin time
08/11/2018 12:00:00 AM

End time
08/17/2018 11:59:59 PM

sample length
15

Usage_thb
76675.4

Ft_thb
3755.4

Max demand_thb
25272.8

Service fee_thb
312.24

Total_thb
113437

Tax_thb
7421.11

(a)

Begin time
10/13/2018 12:00:00 AM

End time
06/26/2019 02:51:54 PM

error out 10
status ✓ code 0
source

(b)

Fig. 24 Labview program UI **a** Calculation of billing cost based on active power **b** Moving average of active power

Comparison to other datasets

Table 5 compares the dataset presented in this paper with related ones. Most measured power consumption in the context of households in several aspects. For example, in the aspect of appliance usage, and room usage (Chavat et al. 2022; Dua and Taniskidou 2017; Candanedo et al. 2017). Some integrates the other context and environment data such as temperature, humidity, billing, etc. (Makonin et al. 2016; Candanedo et al. 2017). In the household context, all used meters data as a main data source while data from sensors were supplementary. In the data center context, CPU/GPU, memory, network power

Table 5 Comparison of existing data sets for data centers

Data set	Domain	Duration	#attr	Detailed description
Dua and Taniskidou (2017)	Household	47 months	9	The dataset was collected from three room types and three submeters
Candanedo et al. (2017)	Household	4.5 months	5	There are 15 rooms in the house. There were main 5 variables: power meters, humidity, wind speed, date time, temperature
Makonin et al. (2016)	Household	24 months	6	The meters were installed for each room in the house. The hourly weather data were also collected. The dataset contains CSV file of electric meter data, weather data, historic electric, water, gas billing and usage
Chavat et al. (2022)	Household	539.2days	9	The dataset contains power usage for 8 appliances. Each appliance was measured with different meters. The total power consumption meter was measured with nine attributes
Bazurto et al (nd)	Server	245 days	15	The dataset contains power consumptions of GPU, CPU, RAM, CPU temperature, etc. for one machine
Ours Chantrapornchai and Chatlatanakulchai (2023)	Data center	20 months	56	The dataset contains the power consumption of each meter divided by area. Each area contains servers and related IT devices

usage were considered Bazurto et al. (nd) while our dataset considered the power usage of data center based on area containing servers and IT equipment.

Future research

The future research can investigate along with the data and related directions.

1. One may consider integrating with other data sources such as a renewable energy to monitor the data generation. Or to integrate with a temperature dataset, using Weather Data API (OpenWeatherMap 2024; TMD 2024). The visualization can show the linkage these new data sources.
2. In the area of exploring the impact of IT workload management on energy consumption, one can install the profiling tools for various workload collection. Since in our real operating environment, the servers cannot be stopped and inserting the logging scripting. We have preliminary tried the tools to the server and peripherals usage and power consumption logging, eg. using RAPHL [34], to collect power consumption of CPU, memory usage, etc, using NVIDIA-SMI (NVIDIA 2024) to collect power consumption of GPU, GPU memory usage, etc. These data may be combined with the aggregate energy data for further analysis. Based on these server load data, the implementation of server provisioning such as the work in Hübotter (2021), Thein et al. (2020) is also possible.
3. The dataset collected can lead to the other research areas with various methods such as optimizing cooling systems based on predictive models (Afroz et al. 2022; Lei and Shao 2023; Bamdad et al. 2023).

Acknowledgements
Authors would like to thank for the support from EPPO, Ministry of Energy, Thailand, and Faculty of Engineering for the project funding. Also, we would like to thank Office of Computing Service (OCS), Kasetsart University for supporting the meter installation process.

Author contributions
W.C. dedicated to to the design and implementation of hardware installation plan as well as programming involving hardware data logger. C.C. contributed to the overall research logistics, python programming, data visualization as software examples, as well as preparing the manuscript.

Funding
Authors would like to thank for the support from EPPO Grant Number 33-4/2560, Ministry of Energy, Thailand, and Faculty of Engineering for the project funding.

Data availability
The python source coAQ3de for required libraries of statsmodel, numpy, sklearn, matplotlib, pmdarima, joblib. The sample 3-month data and source code are also available at https://github.com/cchantra/energydata/blob/master/power_test.csv.gz. The whole source code is also available at the author Github https://github.com/cchantra/energydata.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 3 March 2024 Accepted: 4 April 2024
Published online: 17 April 2024

References

Afroz Z, Shafiullah GM, Urmee T, Shoeb MA, Higgins G (2022) Predictive modelling and optimization of hvac systems using neural network and particle swarm optimization algorithm. Build Environ 209:108681. https://doi.org/10.1016/j.buildenv.2021.108681
Bamdad K, Mohammadzadeh N, Cholette M, Perera S (2023) Model predictive control for energy optimization of hvac systems using energyplus and aco algorithm. Buildings. 13(12):3084. https://doi.org/10.3390/buildings13123084
Bazurto A, Torres D, Asanza V, Estrada R. Data server energy consumption dataset. https://doi.org/10.21227/x6jw-m015
Candanedo LM, Feldheim V, Deramaix D (2017) Data driven prediction models of energy use of appliances in a low-energy house. Energy Build 140:81–97
Chantrapornchai C, Chatlatanakulchai W (2023) Energy meter data of data center in a University. Science Data Bank. https://doi.org/10.57760/sciencedb.10792
Chavat J, Nesmachnow S, Graneri J, Alvez G (2022) ECD-UY, detailed household electricity consumption dataset of Uruguay. Sci Data 9(21):21
DeepMind: DeepMind AI Reduces Google Data Centre Cooling Bill by 40%. https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/. Accessed 8 Oct 2018
Donovan P (2018) The State of AI and Machine Learning in Data Centers. Accessed 19 Sep 2018
Dua D, Taniskidou EK (2017) UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
Fukumoto K, Tamura N, Ishibashi H (2010) Optimizing of it load and facility energy in data center. FUJISU Sci. Tech 46(4):376–382
Gao J (2018) Machine learning applications for data center optimization. Accessed 18 Sep 2018
GrafanaLab: Grafana. Accessed 28 February
Hemsoth N (2016) Mission Possible— Greening the HPC Data Center. Accessed 23 Jan 2016
Hübotter J (2021) Implementation of Algorithms for Right-Sizing Data Centers. doi:1048550/ARXIV.2108.09489
Hyndman RJ, Athanasopoulos G (2018) Forecasting: Principles and Practice. OTexts.com/fpp2, Melborune, Australia
Hyndman RJ, Athanasopoulos G (2020) Forecasing: Principles and Practices, 3rd edn. OText. Chap. Chapter 8 ARIMA models
Influxdata: InfluxDB. It's About Time. https://www.influxdata.com/. Accessed: 8 Feb 2024 (2024)
INTELopensource.org: RUNNING AVERAGE POWER LIMIT — RAPL. Accessed 17 Oct 2018
Jennings C (January 2016) Data Center Energy Consumption. Accessed 23
Lei L, Shao S (2023) Prediction model of the large commercial building cooling loads based on rough set and deep extreme learning machine. J Build Eng 80:107958. https://doi.org/10.1016/j.jobe.2023.107958
Li C, Ding Z, Zhao D, Yi J, Zhang G (2017) Building energy consumption prediction: an extreme deep learning approach. Energies 10(1525):20
Loeffler C (2016) 10 Ways to save energy in your data center. Accessed 25 Jan 2016
Makonin S, Ellert B, Bajic IV, Popowich F (2016) Electricity, water, and natural gas consumption of a residential house in canada from 2012 to 2014. Sci Data 3(160037):1–12

Michael S, Aaron B, Shanti P (2011) NREL RSF Measured Data . https://doi.org/10.25984/1845288. https://data.openei.org/submissions/358

NVIDIA: NVIDIA System Management Interface. Accessed 28 Feb 2024

Office of energy efficiency and renewable energy: energy-efficient cooling control systems for data centers. Accessed 26 Jan 2016

OpenWeatherMap: Weather API. Accessed 28 Feb 2024

Pandas: pandas.DataFrame.bfill. Accessed 1 Mar 2024

Pandas: pandas.DataFrame.corr. Accessed 18 Feb 2024

PEA: Estimate Electric Bill. Accessed 28 Feb 2024

Phelps W (2018) Airflow management for an efficient data center. Accessed 19 Sep 2018

Shoukourian H, Wilde T, Detlef Labrenz AB (eds.) (2017) Proceedings of IEEE International Parallel and Distributed Processing Symposium Workshops

sklearn: sklearn.linear_model.LinearRegression. Accessed 1 Mar 2024

Strom D (2016) Hyperscale data center means different hardware needs, roles for IT. Accessed 25 Jan 2016

statsmodels: Autoregressions. Accessed 26 (February 2024)

Sverdlik Y (2016) Here's how much energy all US Data Centers Consume. Accessed 23 Jan 2016

Thein T, Myo MM, Parvin S, Gawanmeh A (2020) Reinforcement learning based methodology for energy-efficient resource allocation in cloud data centers. J King Saud Univ Comput Inf Sci 32(10):1127–1139. https://doi.org/10.1016/j.jksuci.2018.11.005

TMD: Open Data of TMD. Accessed 28 Feb 2024

UCI: Household Electric Power Consumption. https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set

Zorba S (2024) Modernizing energy systems can reduce primary energy consumption in heating and cooling by up to 50% – UN Report. Accessed 25 January

Publisher's Note