A Systematic Approach for Exploration, Behavior Analysis, and Visualization of Redundant Data Anomalies in Smart Home Energy Consumption Dataset

K. Purna Prakash[*], Y. V. Pavan Kumar[**‡]

[*]School of Computer Science and Engineering, VIT-AP University, Amaravati-522237, Andhra Pradesh, INDIA

**School of Electronics Engineering, VIT-AP University, Amaravati-522237, Andhra Pradesh, INDIA

(kasaraneni.19phd7020@vitap.ac.in, pavankumar.yv@vitap.ac.in)

[‡]Corresponding Author; Y. V. Pavan Kumar, VIT-AP University, Amaravati-522237, Andhra Pradesh, INDIA

Tel: +91863-2370155, pavankumar.yv@vitap.ac.in

Received: 01.12.2021 Accepted: 27.12.221

Abstract - The increase of smart home culture for improved efficiency and comfort in the present energy sector requires paying much attention to big data analytics. Here, the data refers to the energy consumption readings that are continuously captured through smart meters and transmitted to the central computing centres. The entire analysis and decision making in such cases depend on the availability of quality data. However, this data often contains anomalies such as redundancy (duplicated data), which affects their quality. Thus, a systematic approach with three steps (exploration, behavior analysis, and visualization) is proposed in this paper to precisely analyze the redundant data anomalies and their behavior. In exploration, the identification and quantification of redundant data anomalies will be done for all appliances for all available days. This provides the information of the highest and lowest counts of redundancies for all appliances. In behavior analysis, the behavior of redundant data anomalies during various parts of a day will be analysed. The visualization finds the occurrence of redundant data anomalies at the day/hour/minute level. Altogether, these three steps provide a comprehensive analysis of redundant data anomaly behavior that is present in the smart home energy consumption dataset. For the analysis, this paper considers a real-time smart home dataset 'Tracebase'. From this dataset, the appliance 'WaterKettle' is used as an example for the proposed analysis as it exhibits the highest redundancy count when compared to all other appliances. Form the implementation of the proposed approach, it is revealed that there is a high occurrence of redundancy during Daylight hours and is visualized.

Keywords Behavior analysis; Data analysis; Energy consumption data; Redundant data anomaly; Smart home; Visualization.

1. Introduction

Across the world, the conventional power grids are in transition towards smart grids. This transition is not only limited to macrogrids but also applicable to microgrids such as homes, buildings, and cities. The smartness can be embedded into homes by using the equipment viz., sensors, advanced metering infrastructure, computer-operated and controlled power networks, etc. Further, it enables demand-side management, automatic restoration from power supply faults and blackouts, effective consumer interaction and involvement, precise billing, etc. A home is referred to as a smart home whenever it is furnished with the abovementioned equipment and features. The smart home facilitates the integration of renewable energy sources and electric vehicles for efficient demand-side management. This enables the smart meters to effectively communicate with the smart appliances of smart homes connected in the power network to collect the energy consumption data continuously at a predefined rate. Thus, the identification of all the

possible anomalies in this energy consumption data is significant to perform better analytics.

Over the years, technology advancements facilitated the transformation of regular homes into smart homes and enabled the consumers to fulfil all their needs intended for convenient and comfortable living [1]. On the other hand, the availability of high quality (anomaly-free) data becomes a major constraint for information processing and decision making. To learn the technology advancements in anomaly detection, a thorough review of the existing frameworks that relied on artificial intelligence (AI) was discussed in [2]. Big data analytics is one of the technological advancements that provide better insights to energy consumption data. Further, it helps to know the consumers' energy consumption pattern [3]-[4]. Besides, the techniques and applications of data analytics in the domains of smart grids and electricity are thoroughly discussed in [5]. A comprehensive model which provides the details of the key elements and characteristics of the microgrid was discussed in [6]. Further, to overcome the issue of communication failure in the power networks a protocol named hot standby router protocol was presented [7]. Several machine learning algorithms were discussed to thoroughly monitor the non-intrusive load monitoring in the electrical networks [8]. Further, a review was conducted to understand the challenges and potentiality of data analytics in the power systems and future electrical grids [9]-[10].

The energy consumption big data that are streaming in the power network is collected by the smart meter, which is an integral part of the advanced metering infrastructure (AMI) [11]. Moreover, to better understand the energy utilization of one million houses, the smart meter data are synthesized and studied using effective visualizations [12]. With the ramping up of data analytics, especially in the smart grid's context, it is essential to learn the up-to-date knowledge and facets of smart grid data analytics in various subdomains [13]. Generally, in a smart home, the occurrence of anomalies is not only due to the malfunctioning of power networks but also from user operations and behavior [14]. These anomalies and faults are the abnormalities observed in smart grid operations. These faults may arise in physical/ hardware-software/communication levels of the smart grid system [15]. Further, a broad survey was conducted to know the current trends and new perceptions in smart grid power systems [16]. For effective data analytics in smart metering systems, a multi-tier architecture was given [17]. The impact of smart grid technologies on conventional grids was discussed in [18]. The influence of renewable energy resources on a smart grid system was presented [19].

Usually, the appliances of smart buildings generate voluminous data. The proper storage, processing, and analysis of this big data are essential. To realize better decision-making and timely actions, advanced machine learning techniques are needed [20]. Moreover, the infant phase of big data showcases the importance of addressing the issues associated with data acquisition, management, and analysis in smart grids [21]. A thorough review was conducted to learn the threats and vulnerabilities involved in the smart grid [22]. The challenges such as combining big data with electrical systems, complexity in processing, and providing security to big data were mentioned in [23], and a thorough review of big data analytics to deal with such challenges was discussed. The issues related to the load profiles and load forecasting was discussed in [24]. A system was implemented with cloud technology for effective analytics of smart home data using the internet of things (IoT) [25]. The future research trends and the latest technologies of smart grids were reviewed in the context of computing and communication [26]. A framework was presented to detect anomalies in the multivariate smart meter data [27]. A review was performed on applications of big data analytics in the smart grids context [28]. In addition, the key barriers that usually arise in smart grid communication networks and the leading-edge solutions to cope up with these barriers were discussed in [29]. A thorough survey was conducted on the intricate challenges and solutions involved in blending the electrical vehicles, renewable energy sources and demand-side initiatives [30]. Further, a platform for big building data processing was introduced [31]. The big data challenges concerning storage, management, processing, security, and visualization were discussed in [32]. The interoperability of the infrastructure, processing and representation, intelligence in real-time, privacy, and security are also being the major challenges in big data analytics [33]. An elaborate survey was conducted on the critical issues that exist in smart grids related to control and automation technologies, energy storage technologies, power electronics, measurement, information and communication technologies, and sensing [34]. A review was given on the challenges and suitable solutions for big data analytics and smart grids [35].

The above works represent the importance of data analytics and data anomalies in the smart grid environment and various approaches for detecting anomalies in smart home energy consumption data. The cause for such anomalies is that this equipment may send energy consumption data multiple times to the advanced metering infrastructure. Generally, it happens due to the delay in communication networks, which leads to improper acknowledgement from the advanced metering infrastructure, whether the data are received or not. The existence of redundancies affects the integrity of data and leads to improper analytics. So, there is a dire necessity for the accurate identification and analysis of such anomalies. Hence, a systematic approach is proposed in this paper for exploring the behavior of all the possible redundant data anomalies in smart home energy consumption data. Initially, it identifies and quantifies the redundant data that exists in the dataset and further observes the occurrence of redundancy behavior during various parts of the day.

2. Implementation of the Proposed Approach

The proposed approach has three steps, viz., exploration, behavior analysis, and visualization. The conceptual model of the proposed approach is shown in Fig. 1. The smart home energy consumption dataset is considered as an input for the proposed approach. In exploration, the identification and quantification of redundant data anomalies that exist in the smart home energy consumption dataset will be done. This step will be performed on all appliances on each day in

which the appliance is connected. In behavior analysis, the behavior of redundant data anomalies that occurs during various parts of the day will be analyzed.

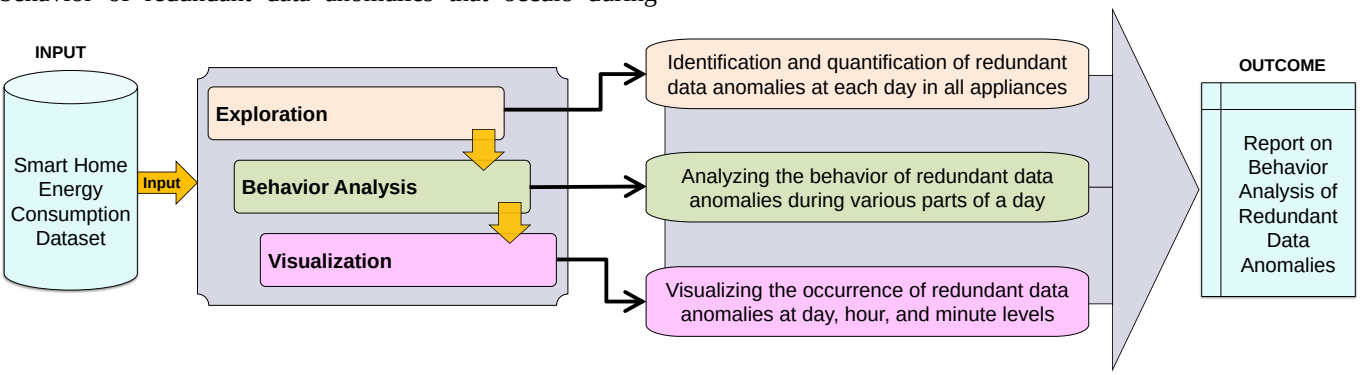


Fig. 1. Conceptual model of the proposed approach.

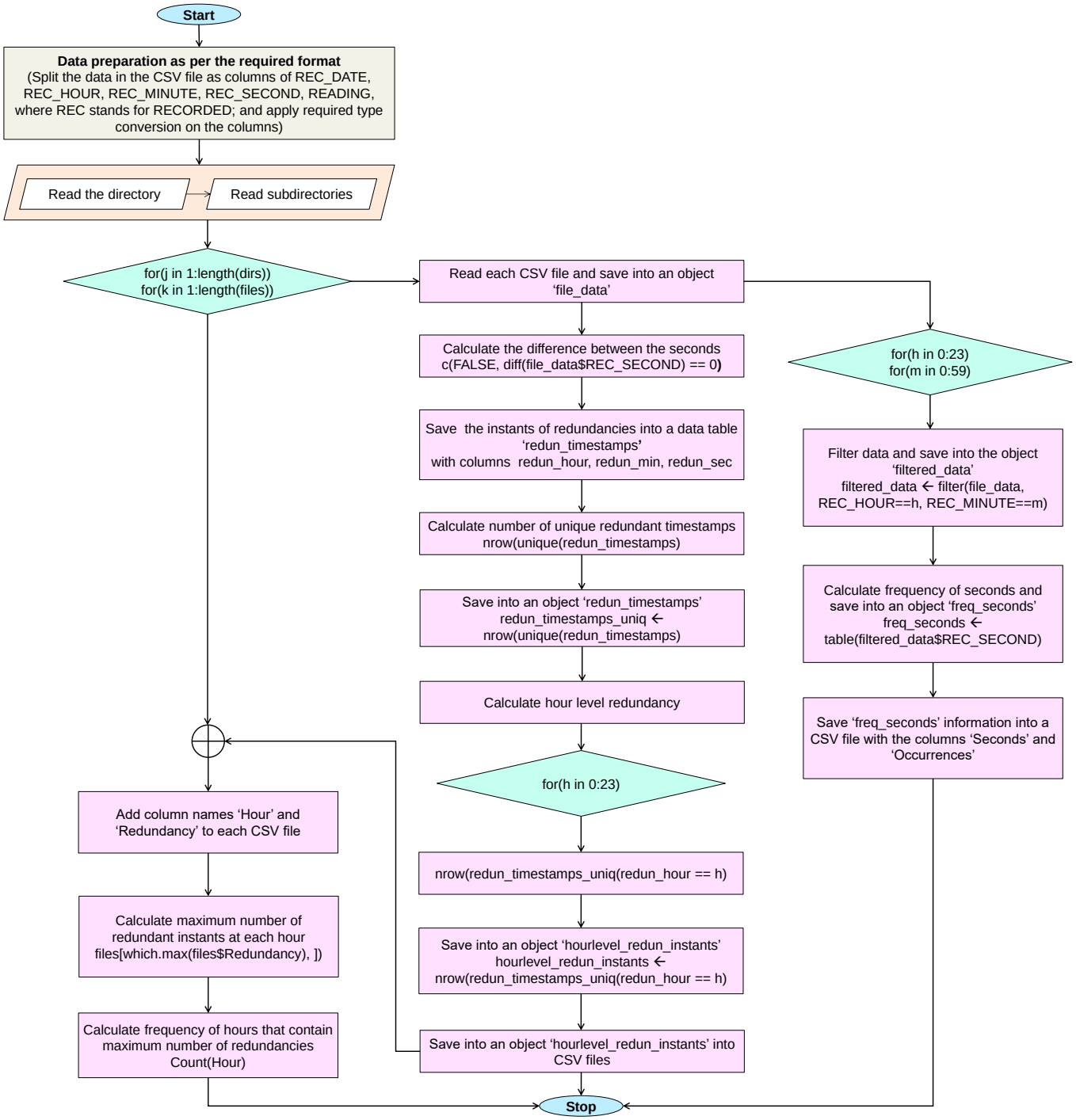


Fig. 2. Implementation steps of the proposed approach.

In visualization, the occurrence of redundant data anomalies will be visualized at day, hour, and minute levels. Altogether, these three steps provide an analysis report as an outcome on the behavior of redundant data anomalies. The implementation steps of the proposed approach are shown in Fig. 2. To execute this approach, a real-time dataset 'Tracebase' is used. This dataset consists of 43 subdirectories in a directory named 'complete' [36]. Each subdirectory represents an appliance and it comprises the timestamped energy consumption readings of the corresponding devices of that appliance. The information of these energy consumption readings is available in comma-separated values (CSV) files, where, each CSV file represents a day that consists of the energy consumption data in string format in a single column. This raw format is not directly useful to perform the analysis. Hence, proper preparation of the dataset is essential to implement any kind of analysis.

For data preparation, the pre-existing single column will be split into multiple columns such as REC_DATE, REC_HOUR, REC_MINUTE, REC_SECOND, and the READING as required for the proposed analysis. Here, the word REC stands for RECORDED. Further, the type conversion on the columns will be applied. Once the dataset is ready, the reading of the directory and subdirectories can

be done. This reading continues till the end of the directory and subdirectories using for(j in 1:length(dirs)) and for(k in 1:length(files)), where 'dirs' represents the directory and 'files' represents the CSV files in the subdirectories. Read each CSV file and save it into an object 'file_data'. To identify the redundant timestamps, the difference between seconds in the column REC_SECOND will be calculated using the function c(FALSE, diff(file_data$REC_SECOND) ==0). Here, FALSE indicates that the process starts from the second record. The redundant timestamps information will be saved into a data table 'redun_timestamps'. This data table consists of the columns 'redun_hour', 'redun_min', and 'redun_sec'. The unique number of redundant timestamps will be calculated and saved using redun_timestamps_uniq ← nrow(unique(redun_timestamps)). This information will be useful for exploring redundant data anomalies. The process continues with the calculation of hour level redundancy in redun_timestamps_uniq.

The redundancies occurred at each hour for(h in 0:23) will be calculated and saved using hourlevel_redun_instants ← nrow(redun_timestamps_uniq(redun_hour==h). This hour level information will be saved into CSV files with the columns 'Hour' and 'Redundancy'. Each CSV file will be read by using for(k in 1:length(files)) and the hour with maximum redundancy can be calculated by using the function files[which.max(files$Redundancy),]. From this, the frequency of hours with maximum redundancy will be calculated using count(Hour). This information will be useful for observing the occurrence of redundancy during the parts of the day. To calculate the occurrence of redundancies, the object 'file_data' will be considered. This data will be filtered by using filter(file_data, REC_HOUR==h, REC_MINUTE==m) in every hour and minute i.e., for(h in 0:23) and for(m in 0:59). This filtered data will be saved into the object named 'filtered_data'. The frequency of seconds in the 'filtered_data' can be calculated by using the function table(filtered_data$REC_SECOND). This frequency statistic will be saved into an object 'freq_seconds' and then into a CSV file with the columns 'Seconds' and 'Occurrences'.

3. Results and Analysis

The simulation results of the proposed approach with respect to its three steps exploration, behavior analysis, and visualization are summarized as follows.

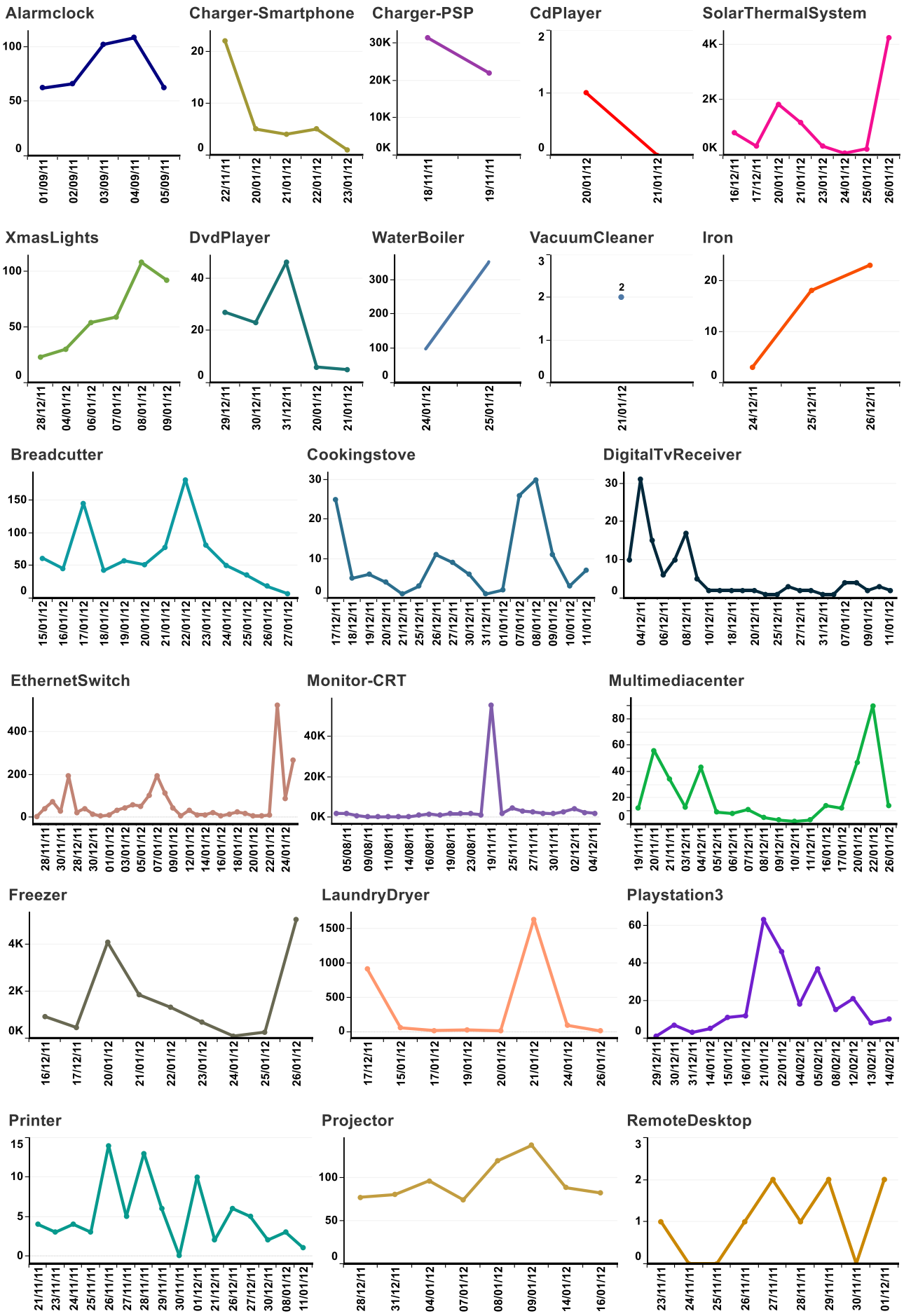
- Exploration of redundant data anomalies day-wise in all appliances is given in section 3.1 using Fig. 3 and Fig. 4.
- Behavior analysis of redundant data anomalies is discussed in section 3.2 using Fig. 5 and Fig. 6.
- Visualization of the occurrence of redundancies in WaterKettle with the device identifier "B81D04" is discussed in section 3.3 using Fig. 7 to Fig. 9.

3.1. Results Pertaining to Exploration

The subplots shown in Fig. 3 represent the variation in the count of redundancies that exist in different appliances of the smart home energy consumption dataset. These subplots are drawn based on the number of days an appliance is connected (on the x-axis) and the count of redundancies (on the y-axis). The quantification of redundancies for each appliance that are observed from Fig. 3 is given as follows.

Alarmclock is connected for 5 days and the highest redundancy count (108) is observed on 04/09/2011, lowest redundancy count (62) is observed on 05/09/2011. Charger-Smartphone is connected for 5 days and the highest redundancy count (22) is observed on 22/11/2011, lowest redundancy count (1) is observed on 23/01/2012. Charger-PSP is connected for 2 days and the highest redundancy count (31434) is observed on 18/11/2011, lowest redundancy count (21941) is observed on 19/11/2011. CdPlayer is connected for 2 days and the highest redundancy count (1) is observed on 20/01/2012, lowest redundancy count (0) is observed on 21/01/2012. SolarThermalSystem is connected for 8 days and the highest redundancy count (4234) is observed on 26/01/2012, lowest redundancy count (64) is observed on 24/01/2012.

XmasLights is connected for 6 days and the highest redundancy count (108) is seen on 08/01/2012, lowest redundancy count (23) is seen on 28/12/2011. DvdPlayer is connected for 5 days and the highest redundancy count (46) is observed on 31/12/2011, lowest redundancy count (5) is observed on 21/01/2012. WaterBoiler is connected for 2 days and the highest redundancy count (350) is observed on 25/01/2012, lowest redundancy count (98) is observed on 24/01/2012. VacuumCleaner is connected for 1 day and the redundancy count (2) is observed on 21/01/2012. Iron is connected for 3 days and the highest redundancy count (23) is observed on 26/12/2011, lowest redundancy count (3) is observed on 24/12/2011. Breadcutter is connected for 13 days and the highest redundancy count (181) is seen on 22/01/2012, lowest redundancy count (7) is seen on 27/01/2012. Cookingstove is connected for 16 days and the highest redundancy count (30) is seen on 08/01/2012, lowest redundancy count (1) is seen on 21/12/2011 and 31/12/2011.


Alarmclock
Charger-Smartphone
Charger-PSP
CdPlayer
SolarThermalSystem
XmasLights
DvdPlayer
WaterBoiler
VacuumCleaner
Iron
Breadcutter
Cookingstove
DigitalTvReceiver
EthernetSwitch
Monitor-CRT
Multimediacenter
Freezer
LaundryDryer
Playstation3
Printer
Projector
RemoteDesktop

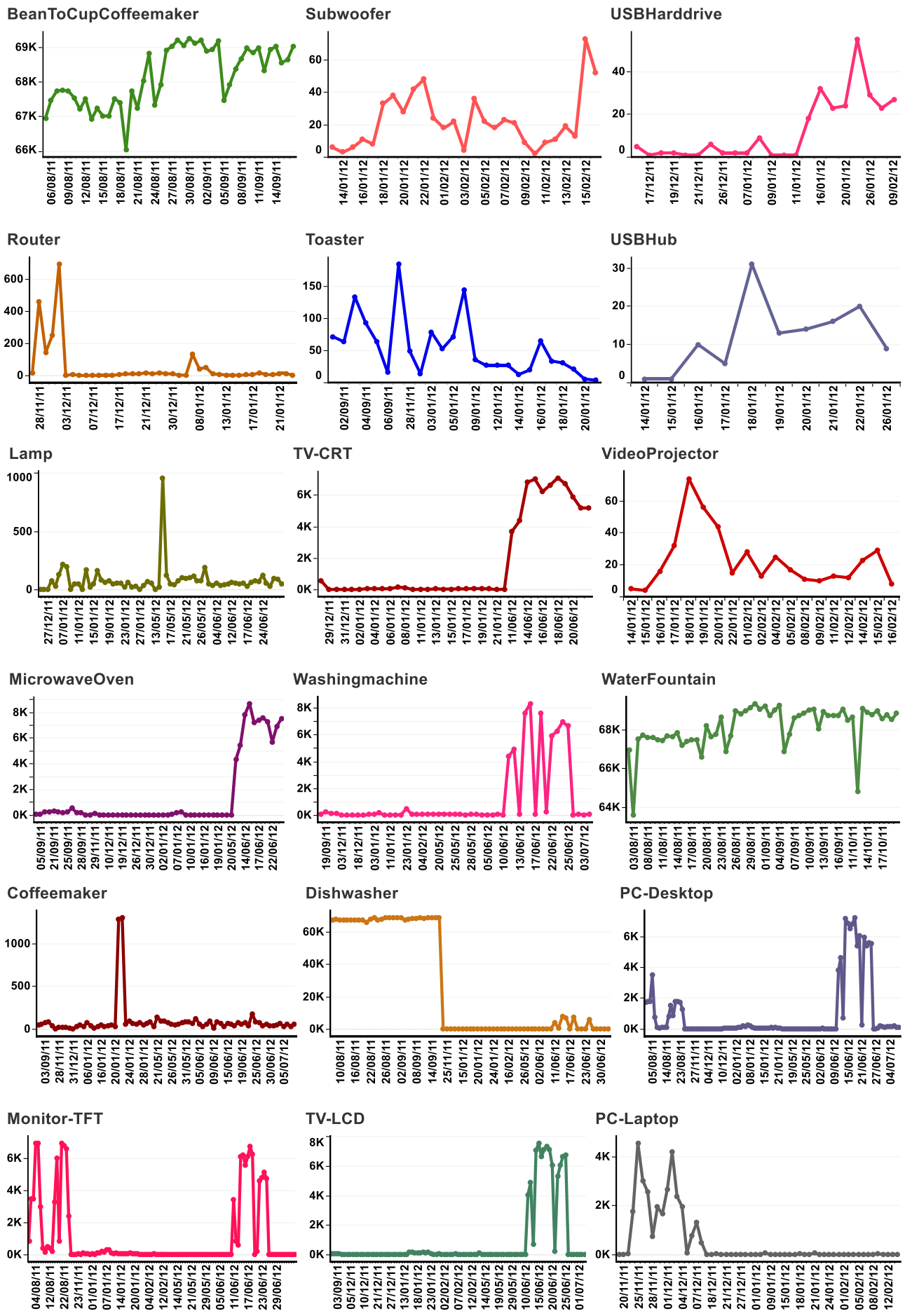

BeanToCupCoffeemaker
Subwoofer
USBHarddrive
Router
Toaster
USBHub
Lamp
TV-CRT
VideoProjector
MicrowaveOven
Washingmachine
WaterFountain
Coffeemaker
Dishwasher
PC-Desktop
Monitor-TFT
TV-LCD
PC-Laptop

Fig. 3. Quantification of identified redundant data anomalies in different appliances (x-axis represents the date and y-axis represents the count of redundancies corresponding to that date in all the subplots).

DigitalTvReceiver is ON for 24 days. Peak redundancy count (31) is seen on 4/12/2011, lowest redundancy count (1) is seen on 21/12/2011, 25/12/2011, 31/12/2011, and 1/1/2012. EthernetSwitch is ON for 33 days and the highest redundancy count (524) is seen on 23/1/2012, lowest redundancy count (1) is seen on 23/11/2011. Monitor-CRT is ON for 26 days and the highest redundancy count (55642) is seen on 19/11/2011, lowest redundancy count (78) is observed on 10/08/2011. Multimediacenter is ON for 17 days and the highest redundancy count (90) is seen on 22/01/2012, lowest redundancy count (2) is seen on 10/12/2011. Freezer is ON for 9 days, whose highest (5060) and lowest (89) redundancy count is seen on 26/1/2012 and 24/1/2012. LaundryDryer is ON for 9 days, in which the highest (1638) and lowest (15) redundancy count is seen on 21/1/2012 and 26/1/2012. Playstation3 is ON for 14 days, in which the highest (63) and lowest (1) redundancy count is seen on 21/1/2012 and 29/12/2011. Printer is ON for 16 days, in which the highest (14) and lowest (0) redundancy count is seen on 26/11/2011 and 30/11/2011. Projector is ON for 8 days, in which the highest (137) and lowest (74) redundancy count is seen on 9/1/2012 and 7/1/2012. RemoteDesktop is ON for 9 days. The highest redundancy count (2) is seen on 27/11/2011, 29/11/2011, and 1/12/2011, lowest redundancy count (0) on 24/11/2011, 25/11/2011, and 30/11/2011.

BeanToCupCoffeemaker is ON for 44 days and the highest redundancy count (69275) is seen on 30/08/2011, lowest redundancy count (66037) is seen on 19/08/2011. Subwoofer is ON for 28 days and the highest redundancy count (73) is seen on 15/02/2012, lowest redundancy count (2) is seen on 10/02/2012. USBHarddrive is ON for 30 days and the highest redundancy count (55) is seen on 22/01/2012, lowest redundancy count (1) is seen on 17/12/2011, 20/12/2011, 21/12/2011, and 9/1/2012 to 11/01/2012. Router is ON for 40 days and the highest redundancy count (697) is seen on 1/12/2011, lowest redundancy count (0) is seen on 3/12/2011, 6/12/2011, and 8/12/2011 to 10/12/2011. Toaster is ON for 25 days and the highest redundancy count (185) is seen on 23/09/2011, lowest redundancy count (4) is seen on 21/01/2012. USBHub is ON for 10 days and the highest redundancy count (31) is seen on 18/01/2012, lowest redundancy count (1) is seen on 14/01/12 and 15/01/12. Lamp is ON for 86 days and the highest redundancy count (959) is seen on 15/05/2012, lowest redundancy count (1) is seen on 26/12/2011 and 12/01/2012. TV-CRT is ON for 36 days and the highest redundancy count (7106) is seen on 18/06/2012, lowest redundancy count (16) is seen on 09/01/2012. VideoProjector is ON for 19 days and the highest redundancy count (74) is seen on 18/01/2012, lowest redundancy count (4) is seen on 15/1/2012. MicrowaveOven is ON for 60 days, for which the highest (8695) and lowest (1) redundancy count is seen on 15/06/2012 and 09/12/2011.

Washingmachine is ON for 56 days, in which the highest (8352) and lowest (2) redundancy count is seen on 15/6/2012 and 10/12/2011. WaterFountain is ON for 56 days and the highest redundancy count (69358) is seen on 30/08/2011, lowest redundancy count (63598) is seen on 3/8/2011. Coffeemaker is ON for 82 days and the highest redundancy count (1309) is seen on 22/01/2012, lowest redundancy count (0) is seen on 31/12/2011. Dishwasher is ON for 76 days and the highest redundancy count (69102) is seen on 31/8/2011, lowest redundancy count (22) is seen on 3/2/2012. PC-Desktop is ON for 151 days and the highest redundancy count (7238) is seen on 18/6/2012, lowest redundancy count (0) is seen on 9/12/2011 and 10/12/2011. Monitor-TFT is ON for 190 days and the highest redundancy count (6969) is seen on 19/8/2011, lowest redundancy count (0) is seen on 19/5/2012, 29/5/2012, 31/5/2012, 1/6/2012, 8/6/2012, 9/6/2012, 28/6/12, 30/6/2012, 3/7/2012, 4/7/2012, 7/7/2012, and 8/7/2012. TV-LCD is ON for 119 days, for which the highest (7590) and lowest (0) redundancy count is seen on 15/6/2012 and 14/5/2012. PC-Laptop is ON for 67 days and

the highest redundancy count (4564) is seen on 25/11/2011, lowest redundancy count (1) is seen on 15/1/2012 to 17/1/2012, 24/1/2012, and 25/1/2012. Refrigerator is ON for 206 days and the highest redundancy count (69503) is seen on 4/9/2011, lowest redundancy count (1) is seen on 25/12/2011. Amplifier is ON for 89 days and the highest redundancy count (232) is seen on 5/2/2012, the lowest redundancy count (2) is seen on 25/1/2012. WaterKettle is ON for 134 days, whose highest (82227) and lowest (1) redundancy count is seen on 20/11/2011, and 23/11/2011, 31/12/2011 respectively.

From the above remarks, it is noticed that the appliances CdPlayer, Coffeemaker, Monitor-TFT, PC-Desktop, Printer, RemoteDesktop, Router and TV-LCD are containing 0 (zero) redundancy count on some days. Further, the summary of these observations on redundancy at various dates with their highest and lowest counts are given in Table 1. The highest counts of redundancy in all appliances with the respective devices are shown in Fig. 4. From this, it is evident that the appliance 'WaterKettle' contains the highest redundancy count of 82206. Thus, it is used for further analysis.

Table 1. Summary of redundancy counts at each day in different appliances

S.No.	Appliance	No. of Days Connected	Observation on Highest Redundancy Counts		Observation on Lowest Redundancy Counts	
			Date(s) with Highest Redundancy Instants	Corresponding Redundancy Count	Date(s) with Lowest Redundancy Instants	Corresponding Redundancy Count
1	Alarmclock	5	04/09/2011	108	05/09/2011	62
2	Charger-Smartphone	5	22/11/2011	22	23/01/2012	1
3	Charger-PSP	2	18/11/2011	31434	19/11/2011	21941
4	CdPlayer	2	20/01/2012	1	21/01/2012	0
5	SolarThermalSystem	8	26/01/2012	4234	24/01/2012	64
6	XmasLights	6	08/01/2012	108	28/12/2011	23
7	DvdPlayer	5	31/12/2011	46	21/01/2012	5
8	WaterBoiler	2	25/01/2012	350	24/01/2012	98
9	VacuumCleaner	1	21/01/2012	2	-	-
10	Iron	3	26/12/2011	23	24/12/2011	3
11	Breadcutter	13	22/01/2012	181	27/01/2012	7
12	Cookingstove	16	08/01/2012	30	21/12/2011, 31/12/2011	1
13	DigitalTvReceiver	24	04/12/2011	31	21/12/2011, 5/12/2011, 31/12/2011, 01/01/2012	1
14	EthernetSwitch	33	23/01/2012	524	23/11/2011	1
15	Monitor-CRT	26	19/11/2011	55642	10/08/2011	78
16	Multimediacenter	17	22/01/2012	90	10/12/2011	2
17	Freezer	9	26/01/2012	5060	24/01/2012	89
18	LaundryDryer	9	21/01/2012	1638	26/01/2012	15
19	Playstation3	14	21/01/2012	63	29/12/2011	1
20	Printer	16	26/11/2011	14	30/11/2011	0
21	Projector	8	09/01/2012	137	07/01/2012	74
22	RemoteDesktop	9	27/11/2011, 29/11/2011, 01/12/2011	2	24/11/2011, 5/11/2011, 30/11/2011	0
23	BeanToCupCoffeemaker	44	30/08/2011	69275	19/08/2011	66037
24	Subwoofer	28	15/02/2012	73	10/02/2012	2
25	USBHarddrive	30	22/01/2012	55	17/12/2011, 20/12/2011, 21/12/2011, 09/01/2012, 11/01/2012	1
26	Router	40	01/12/2011	697	03/12/2011, 06/12/2011, 08/12/2011 to 10/12/2011	0
27	Toaster	25	23/09/2011	185	21/01/2012	4
28	USBHub	10	18/01/2012	31	14/01/12, 15/01/12	1
29	Lamp	86	15/05/2012	959	26/12/2011, 12/01/2012	1
30	TV-CRT	36	18/06/2012	7106	09/01/2012	16
31	VideoProjector	19	18/01/2012	74	15/01/2012	4
32	MicrowaveOven	60	15/06/2012	8695	09/12/2011	1
33	Washingmachine	56	15/06/2012	8352	10/12/2011	2
34	WaterFountain	56	30/08/2011	69358	03/08/2011	63598
35	Coffeemaker	82	22/01/2012	1309	31/12/2011	0
36	Dishwasher	76	31/08/2011	69102	03/02/2012	22
37	PC-Desktop	151	18/06/2012	7238	09/12/2011, 10/12/2011	0
38	Monitor-TFT	190	19/08/2011	6969	19/05/2012, 29/05/2012, 31/05/2012, 01/06/2012, 08/06/2012, 09/06/2012, 28/06/2012, 30/06/2012, 03/07/2012, 04/07/2012, 07/07/2012, 08/07/2012	0
39	TV-LCD	119	15/06/2012	7590	14/05/2012	0
40	PC-Laptop	67	25/11/2011	4564	15/01/2012 to 17/01/2012, 24/01/2012, 25/01/2012	1
41	Refrigerator	206	04/09/2011	69503	25/12/2011	1
42	Amplifier	89	05/02/2012	232	25/01/2012	2
43	WaterKettle	134	20/11/2011	82227	23/11/2011, 31/12/2011	1

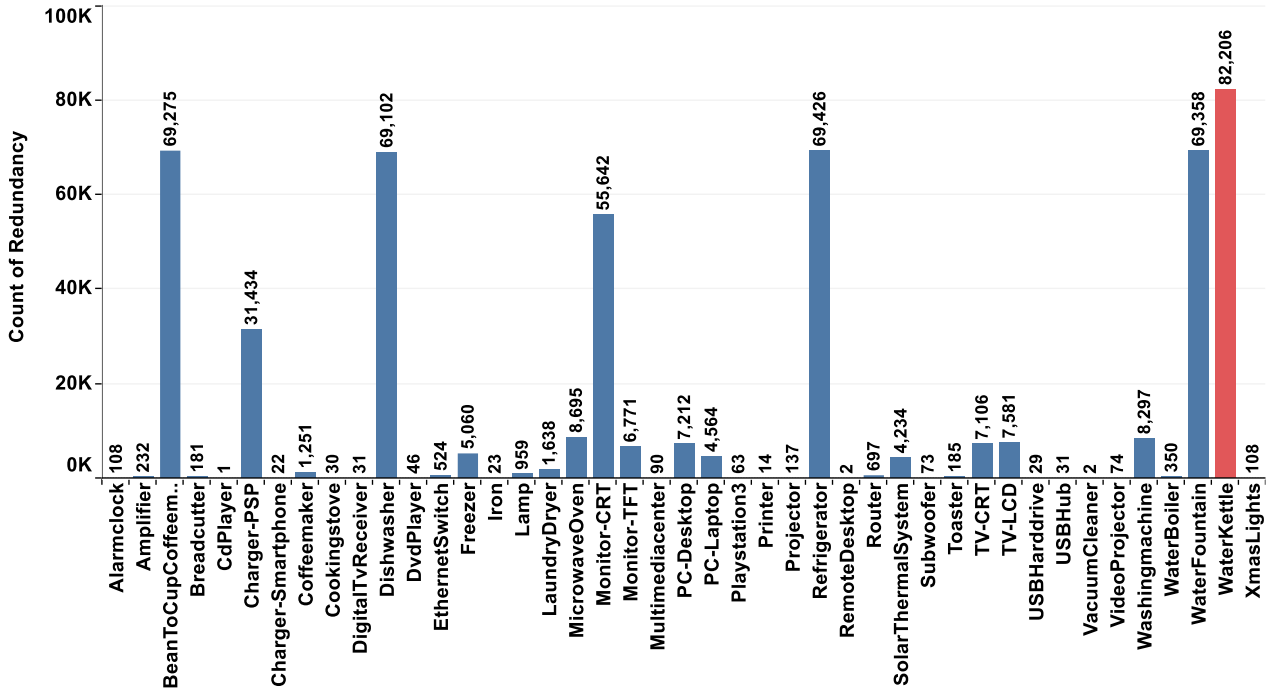


Fig. 4. The highest count of redundancy in all appliances at each device.

3.2. Results Pertaining to Behavior Analysis

There are 134 CSV files in WaterKettle. Each CSV file represents a day, where the WaterKettle is connected for 134 days with different device identifiers. All these 134 CSV files are considered for analyzing the behavior of redundant data in the smart home energy consumption dataset. To accomplish this, the frequency of hours that contain the highest redundancy count is considered as shown in Fig. 5. From this Fig. 5, it is observed that the hours 14, 19, and 21 are having the highest frequency i.e., 10.

The behavior of redundant data in WaterKettle during the parts of the day is shown in Fig. 6. The parts of the day in Darmstadt, Germany, are considered Twilight, Daylight, and Night [37]. The hours of a day are distributed among these parts of the day. The Twilight hours are 7, 20, 21, 22, 23, Daylight hours are 8 to 19, Night hours are 0 to 6. From Fig. 6, it is noticed that frequency of hours with the highest redundancy during Daylight is 77, Twilight is 29, and Night is 28. This analysis reveals the information that the highest redundancy has occurred during the Daylight hours.

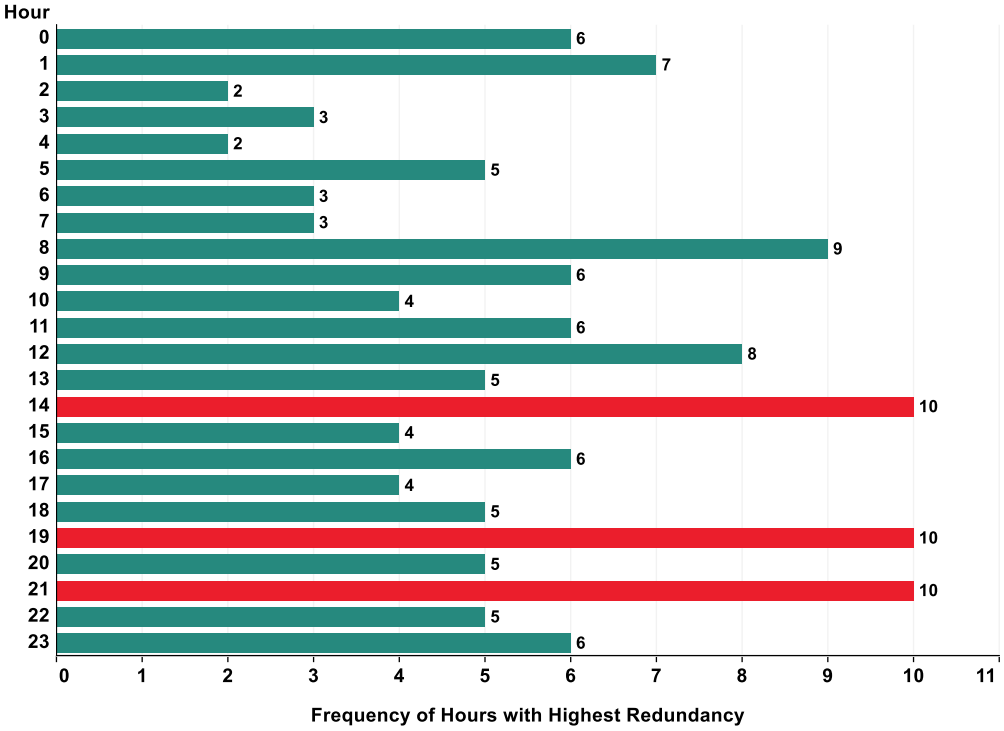


Fig. 5. Frequency of hours with the highest redundancy.

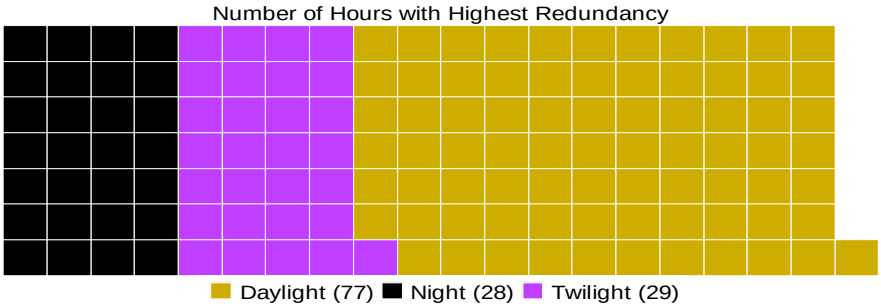


Fig. 6. The behavior of redundant data anomalies during the parts of the day.

3.3. Results Pertaining to Visualization

The highest redundancy is observed on 20/11/2011 in WaterKettle. So, the respective device identifier 'B81D04' is considered from the file 'dev_B81D04_2011.11.20.csv'. The WaterKettle consists of 17 CSV files with this device identifier. All these files are considered for day and hour level analysis of redundancies and are visualized as shown in Fig. 7. The observations of redundancy are given as follows.

The total count of redundancy observed on 18/11/2011 is 38802 and the highest redundancy count (3600) is observed at hours 1 and 23, lowest redundancy count (0) is observed at hours 0, and 9 to 19. The total count of redundancy observed on 19/11/2011 is 77556 and the highest redundancy count (3600) is observed at hours 1 and 3, the lowest redundancy count (0) is observed at hour 0. The total count of redundancy observed on 20/11/2011 is 82206 and the highest redundancy count (3598) is observed at hours 1 and 3, the lowest redundancy count (0) is observed at hour 0. The total count of redundancy observed on 22/11/2011 is 1440 and the highest redundancy count (152) is observed at hour 17, the lowest redundancy count (0) is observed at hours 0, and 10 to 13. The total count of redundancy seen on 24/11/2011 is 1787 and the highest redundancy count (306) is seen at hour 18, lowest redundancy count (0) is seen at hours 10 to 12.

The total count of redundancy observed on 25/11/2011 is 4552 and the highest redundancy count (313) is observed at hour 22, lowest redundancy count (0) is observed at hours 10 to 13. The total count of redundancy observed on 26/11/2011 is 3037 and the highest redundancy count (309) is observed at hour 3, lowest redundancy count (44) is observed at hour 12. The total count of redundancy observed on 27/11/2011 is 2594 and the highest redundancy count (510) is observed at hour 2 and the lowest redundancy count (0) is observed at hour 16. The total count of redundancy observed on 28/11/2011 is 834 and the highest redundancy count (156) is observed at hour 20, lowest redundancy count (0) is observed at hours 9 to 19. The total count of redundancy observed on 29/11/2011 is 1941 and the highest redundancy count (290) is observed at hour 23, lowest redundancy count (0) is observed at hours 14 to 17. The total count of redundancy observed on 01/12/2011 is 2713 and the highest redundancy count (275) is observed at hour 3, lowest redundancy count (0) is observed at hours 0 at hour 14. The total count of redundancy observed on 02/12/2011 is 4199 and the highest redundancy count (654) is observed at hour 5 and the lowest redundancy count (0) is observed at hour 14.

The total count of redundancy observed on 03/12/2011 is 2409 and the highest redundancy count (301) is observed at hour 4, lowest redundancy count (41) is observed at hour 18. The total count of redundancy observed on 04/12/2011 is 1964 and the highest redundancy count (293) is observed at hour 6, lowest redundancy count (11) is observed at hour 4. The total count of redundancy observed on 05/12/2011 is 79 and the highest redundancy count (51) is observed at hour 8, lowest redundancy count (0) is observed at hours 2, 3, 5, 7, and 9 to 23. The total count of redundancy observed on 06/12/2011 is 806 and the highest redundancy count (71) is observed at hour 21, lowest redundancy count (0) is observed at hours 0, 1, 2, 4, 5, and 14. The total count of redundancy observed on 07/12/2011 is 1295 and the highest redundancy count (67) is observed at hours 9, 16, lowest redundancy count (41) is observed at hour 12.

From the day and hour levels redundancy analysis, it is observed that the day 20/11/2011 has the highest redundancy count (3598) at hours 1 and 3. Hence, among these, hour 1 is considered to analyze the occurrence of redundancies in all minutes and seconds. The occurrence of redundancies in all minutes of hour 1 is shown in Fig. 8. From this, it is understood that every minute has some redundancy. The redundancy occurrences 2, 3, 4 is observed in the minutes 1, 2, 4, 7, 8, 9, 12, 14, 16 to 19, 21, 22, 24, 25, 26, 28, 29, 30, 31, 34, 36 to 42, 48 to 54, 56, 58, and 59. The redundancy occurrences 3,4 is observed in minutes 0, 3, 5, 6, 10, 11, 13, 15, 20, 23, 27, 32, 33, 35, 43 to 47, and 57. Finally, the combination of 1, 2, 3, 4 occurrences are observed in minute 55. From this, it is evident that more than fifty per cent of the minutes have the redundancy occurrences of 2, 3, and 4.

The broad picture of the occurrence of redundancies at all seconds of a minute is shown in Fig. 9. For the explanation purpose, minute 10 and minute 55 of hour 1 are considered. The seconds are considered on the x-axis and the occurrence of redundancies is considered on the y-axis. From Fig. 9 (a), it is noticed that there is a minimum of 3 occurrences of redundancy in the considered minute and there are 4 occurrences in the seconds 3, 13, 20, 27, 39, 46, 52, and 58. From Fig. 9 (b), it is observed that the occurrence of redundancies is a combination of 1, 2, 3, and 4. There is only 1 occurrence at second 37, 2 occurrences at seconds 16, 48, and 54, 4 occurrences at seconds 2, 12, 19, 30, 36, 43, 53, and 55. All the remaining seconds have 3 occurrences.

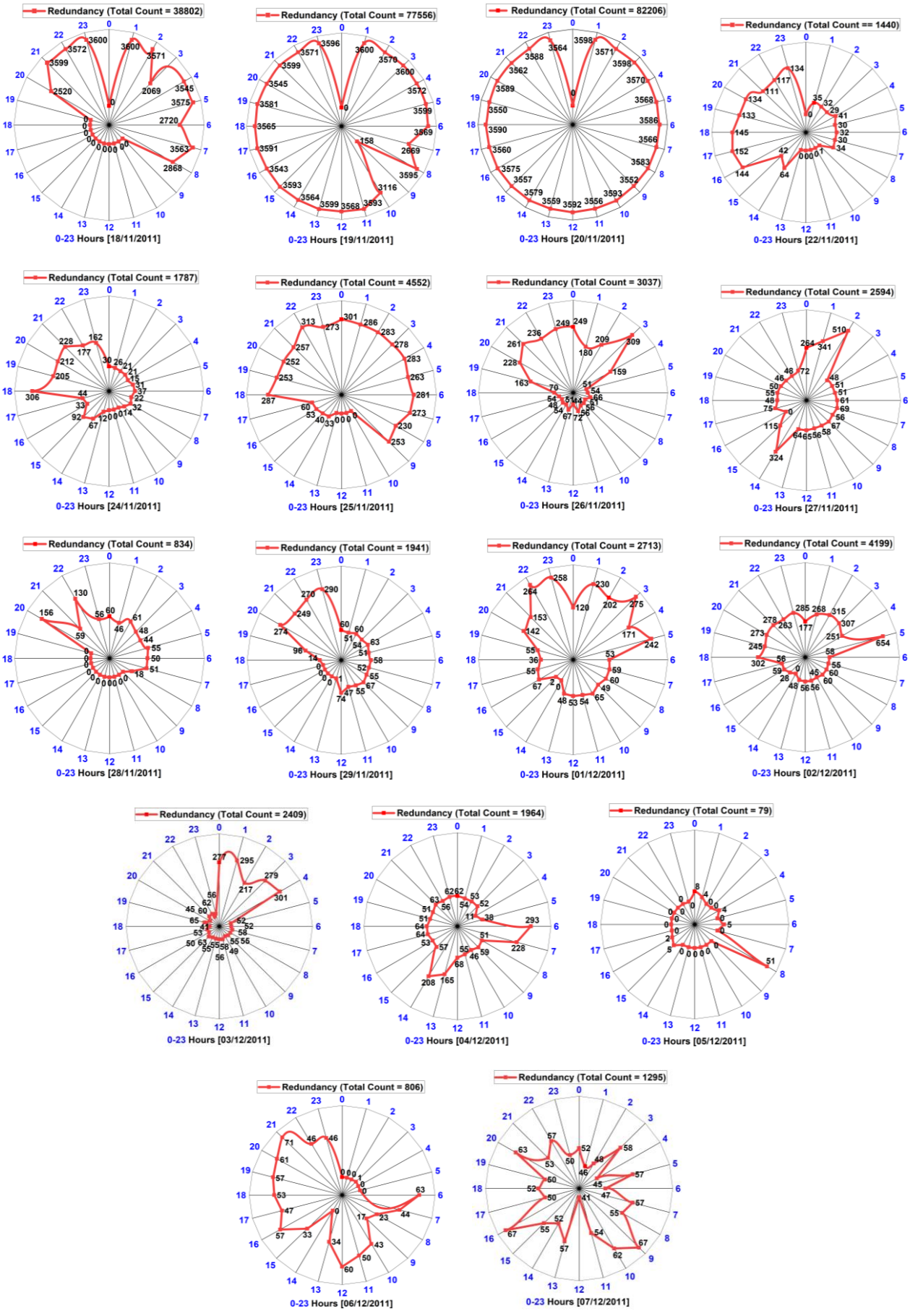


Fig. 7. Redundancy at day and hour levels in WaterKettle.

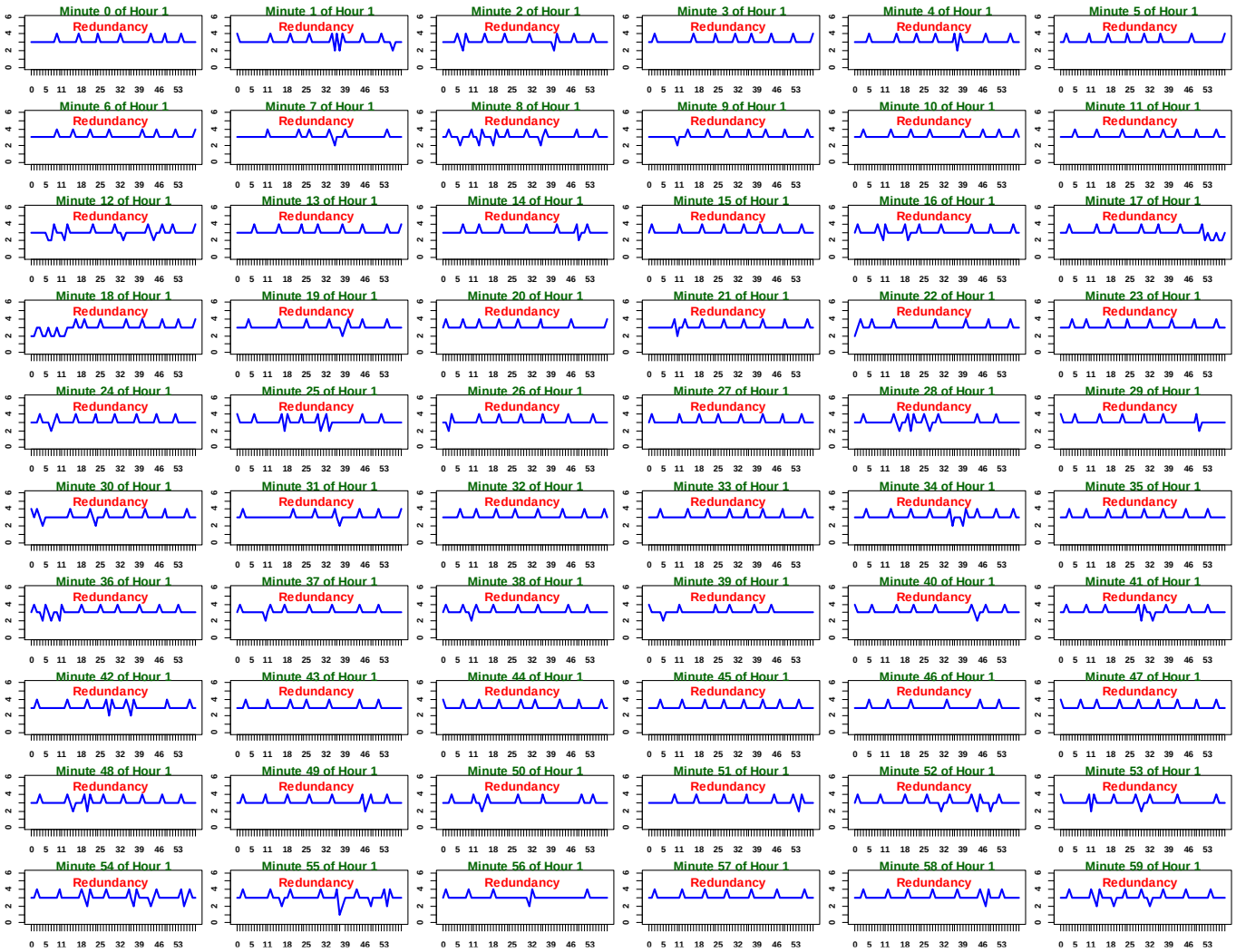
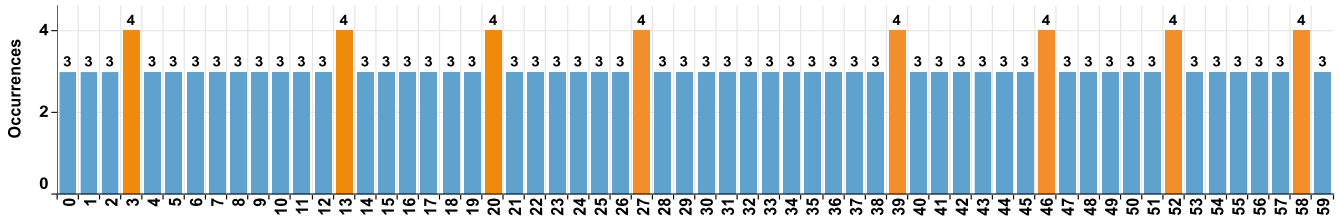
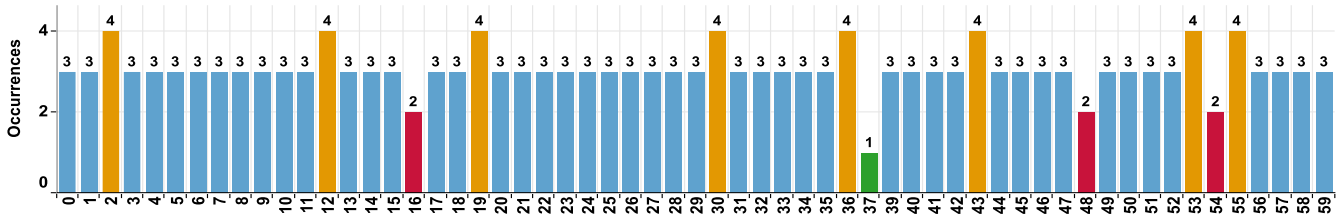


Fig. 8. Occurrence of redundancies on 20/11/2011 at each minute of Hour 1 in WaterKettle.



(a) Occurrences in Hour 1 Minute 10



(b) Occurrences in Hour 1 Minute 55

Fig. 9. Occurrence of redundancies in WaterKettle on 20/11/2011 (x-axis represents the seconds)

4. Conclusions

This paper implements a systematic approach for exploration, behavior analysis, and visualization of redundant data anomalies in the smart home energy consumption dataset. The proposed approach has successfully analyzed all the redundant data anomalies in each day of all the appliances. The salient observations drawn from the implementation of the proposed approach with respect to the considered dataset are given as follows.

- The exploration reveals that the appliance "WaterKettle" consists of the highest redundancies 82227 shared by two

devices (i.e., 82206 in the device "B81D04", 21 in the device "11F01E") on 20/11/2011.

- The behavior analysis reveals that the majority of redundancies occur during Daylight hours.
- The visualization gives a clear picture of the occurrence of redundancies in the device "B81D04" of the WaterKettle.

Hence, the proposed systematic approach in this paper provides a comprehensive report on the presence and behavior of the redundant data anomalies present in energy consumption readings. This helps in the process of data cleaning, which is typically required to perform precise analytics and decision making in smart homes.

Acknowledgements

This work was supported in part by Project Grant No: SRG/2019/000648, sponsored by the Start-up Research Grant (SRG) scheme of the Science and Engineering Research Board (SERB), a statutory body under the Department of Science and Technology (DST), Government of INDIA.

References

[1] A. Zielonka, M. Woźniak, S. Garg, G. Kaddoum, M. J. Piran and G. Muhammad, "Smart homes: how much will they support us? a research on recent trends and advances," IEEE Access, vol. 9, pp. 26388-26419, 2021.

[2] Y. Himeur, G. Khalida, A. Alsalemi, B. Faycal and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," Applied Energy, vol. 287, pp. 116601, 2021.

[3] O. Simona-Vasilica, B. Adela, T. B. George, C. I. Maria and B. M. Alexandru, "Insights into demand-side management with big data analytics in electricity consumers' behaviour," Computers and Electrical Engineering, vol. 89, pp. 106902, 2021.

[4] Z. Niu, W. Junqi, L. Xiufeng, L. Huang and P. N. Sieverts, "Understanding energy demand behaviors through spatio-temporal smart meter data analysis," Energy, vol. 226, pp. 120493, 2021.

[5] D. Syed, Z. Ameema, S. R. Shady and O. Bouhali, "Smart grid big data analytics: survey of technologies, techniques, and applications," IEEE Access, vol. 9, pp. 59564-59585, 2021.

[6] K. S. Rao and Y. V. P. Kumar, "Comprehensive modelling of renewable energy based microgrid for system level control studies," International Journal of Renewable Energy Research, vol. 11, no. 1, pp. 223-234, 2021.

[7] G. P. Reddy and Y. V. P. Kumar, "Retrofitted IoT based communication network with hot standby router protocol and advanced features for smart buildings," International Journal of Renewable Energy Research, vol. 11, no. 3, pp. 1354-1369, 2021.

[8] B. G. Fethi, R. Bayindir, S. Vadi, "Comprehensive non-intrusive load monitoring process: Device event detection, device feature extraction and device identification using KNN, random forest and decision tree," 10th International Conference on Renewable Energy Research and Applications (ICRERA), Istanbul, Turkey, 26-29 Sep., 2021.

[9] F. V. Scheidt, M. Hana, N. Ludwig, R. Bent, P. Staudt and W. Christof, "Data analytics in the electricity sector - a quantitative and qualitative literature review," Energy and AI, vol. 1, pp. 100009, 2020.

[10] K. Mladen, P. Pinson, O. Zoran, S. Grijalva, H. Tao and R. Bessa, "Big data analytics for future electricity grids," Electric Power Systems Research, vol. 189, pp. 106788, 2020.

[11] M. Morteza, A. Ghassemi and T. G. Aaron, "Fast big data analytics for smart meter data," IEEE Open Journal of the Com Soc, vol. 1, pp. 1864-1871, 2020.

[12] G. Ragini, A. R. Al-Ali, Z. A. Imran and D. K. Sajal, "Big data energy management, analytics and visualization for residential areas," IEEE Access, vol. 8, pp. 156153-156164, 2020.

[13] R. Bruno and S. Chren, "Smart grids data analysis: a systematic mapping study," IEEE Transactions on Industrial Informatics, vol. 16, no. 6, pp. 3619-3639, 2020.

[14] M. Yamauchi, O. Yuichi, M. Murata, U. Kensuke and Y. Kato, "Anomaly detection in smart home operation from user behaviors and home conditions," IEEE Transactions on Consumer Electronics, vol. 66, no. 2, pp. 183-192, 2020.

[15] A. E. R. Labrador and A. Taufik, "Faults in smart grid systems: monitoring, detection and classification," Electric Power Systems Research, vol. 189, pp. 106602, 2020.

[16] M. S. Ibrahim, W. Dong and Y. Qiang "Machine learning driven smart electric power systems: Current trends and new perspectives," Applied Energy, vol. 272, pp. 115237, 2020.

[17] J. C. Olivares-Rojas, E. Reyes-Archundia, A. J. Gutiérrez-Gnecchi, J. W. González-Murueta and J. Cerda-Jacobo "A multi-tier architecture for data analytics in smart metering systems," Simulation Modelling Practice and Theory, vol. 102, pp. 102024, 2020.

[18] I. Colak, R. Bayindir and S. Sagiroglu, "The effects of the smart grid system on the national grids," 8th IEEE International Conference on Smart Grid (icSmartGrid), Paris, France, 17-19 June, 2020.

[19] A. Faten, I. Colak, I. Garip and H. I. Bulbul, "Impacts of renewable energy resources in smart grid," 8th IEEE International Conference on Smart Grid (icSmartGrid), Paris, France, 17-19 June, 2020.

[20] B. Qolomany, A. Al-Fuqaha, G. Ajay, B. Driss, S. Alwajidi, Q. Junaid and F. C. Alvis, "Leveraging machine learning and big data for smart buildings: a comprehensive survey," IEEE Access, vol. 7, pp. 90316-90356, 2019.

[21] G. Maedeh, H.D. Sarineh and P. Siano, "Big data issues in smart grids: a survey," IEEE Systems Journal, vol. 13, no. 4, pp. 4158-4168, 2019.

[22] S. Sagiroglu, Y. Canbay and I. Colak "Solutions and suggestions for smart grid threats and vulnerabilities," International Journal of Smart Grid, vol. 9, no. 4, 2053-2063, 2019.

[23] B. P. Bishnu, P. Sumit, Y. Luo, M. Manish, C. Kwok, T. Reinaldo, H. Rob, S.M. Kurt, Z. Rui, P. Zhao, M. Milos, Z. Song, X. Zhang, "Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions," IET Smart Grid, vol. 2, pp. 141-154, 2019.

[24] W. Yi, Q. Chen, H. Tao and K. Chongqing, "Review of smart meter data snalytics: applications, methodologies, and challenges," IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 3125-3148, 2019.

[25] A. Yassine, S. Singh, H. M. Shamim and G. Muhammad, G., "IoT big data analytics for smart homes with fog and cloud computing," Future Generation Computer System, vol. 91, pp. 563–573, 2019.

[26] A. Bani-Ahmed, A. Nasiri and S. Igor, "Foundational support systems of the smart grid: State of the art and future trends," International Journal of Smart Grid, vol. 2, no. 1, pp. 1-12, 2018.

[27] M. Ramin and J. Wang "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," IEEE Transactions on Smart Grid, vol. 9, no. 6, pp. 5820-5830, 2018.

[28] Z. Yang, T. Huang and F. E. Bompard, "Big data analytics in smart grids: a review," Energy Informatics, vol. 1, no. 8, 2018.

[29] M. Yesilbudak and I. Colak, "Main barriers and solution proposals for communication networks and information security in smart grids," 6th IEEE International Conference on Smart Grid (icSmartGrid), Nagasaki, Japan, 4-6 December, 2018.

[30] M. Yesilbudak, A. Colak, "Integration challenges and solutions for renewable energy sources, electric vehicles and demand-side initiatives in smart grids," 7th International Conference on Renewable Energy Research and Applications (ICRERA), Paris, France, 14-17 Oct., 2018.

[31] L. Lucy, D. Vionnet, B. Jean-Philippe and H. Jean, "Big building data – a big data platform for smart buildings," Energy Procedia, vol. 122, pp. 589-594, 2017.

[32] C. Tu, H. Xi, Z. Shuai, and J. Fei, "Big data issues in smart grid - a review," Renewable and Sustainable Energy Reviews, vol. 79, pp. 1099-1107, 2017.

[33] J. Hu and V. V. Athanasios, "Energy big data analytics and security: challenges and opportunities," IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2423-2436, 2016.

[34] I. Colak, S. Sagiroglu, G. Fulli, M. Yesilbudak and C. Catalin-Felix, "A survey on the critical issues in smart grid technologies," Renewable and Sustainable Energy Reviews, vol. 54, pp. 396–405, 2016.

[35] S. Sagiroglu, R. Terzi, Y. Canbay, and I. Colak, "Big data issues in smart grid systems," IEEE International Conference on Renewable Energy Research and Applications (ICRERA), Birmingham, UK, Nov., 2016.

[36] The tracebase power consumption dataset. (http://www.tracebase.org/).

[37] Night, Twilight, and Daylight Times of a Day in Darmstadt, Germany. (https://www.timeanddate.com/astronomy/germany/darmstadt).